

格致方法·定量研究系列　吴晓刚　主编

虚拟变量回归

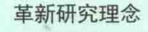
[美] 梅丽莎·A.海蒂（Mellissa A.Hardy）著
贺光烨 译

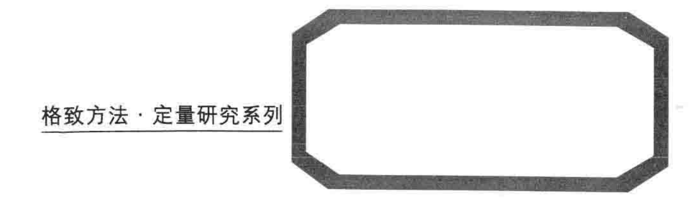
革新研究理念

丰富研究工具

最权威、最前沿的定量研究方法指南

03

格致出版社  上海人民出版社



虚拟变量回归

[美] 梅丽莎·A.海蒂(Mellissa A.Hardy) 著
贺光烨 译

SAGE Publications ,Inc.

格致出版社 上海人民出版社

图书在版编目(CIP)数据

虚拟变量回归/(美)梅丽莎·A.海蒂著;贺光烨译.—上海:格致出版社:上海人民出版社,2016.8
(格致方法·定量研究系列)
ISBN 978-7-5432-2642-5

Ⅰ.①虚… Ⅱ.①梅… ②贺… Ⅲ.①变量-回归分析 Ⅳ.①O174

中国版本图书馆CIP数据核字(2016)第158390号

责任编辑 顾 悦 裴乾坤

格致方法·定量研究系列

虚拟变量回归

[美]梅丽莎·A.海蒂 著
贺光烨 译

出 版 世纪出版股份有限公司 格致出版社
世纪出版集团 上海人民出版社
(200001 上海福建中路193号 www.ewen.co)



编辑部热线 021-63914988
市场部热线 021-63914081
www.hibooks.cn

发 行 上海世纪出版股份有限公司发行中心

印 刷 浙江临安曙光印务有限公司
开 本 920×1168 1/32
印 张 4.5
字 数 88,000
版 次 2016年8月第1版
印 次 2016年8月第1次印刷

ISBN 978-7-5432-2642-5/C·150

定价:25.00元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的"格致方法·定量研究系列"丛书,精选了世界著名的 SAGE 出版社定量社会科学研究丛书,翻译成中文,起初集结成八册,于 2011 年出版。这套丛书自出版以来,受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择,该丛书经过修订和校正,于 2012 年以单行本的形式再次出版发行,共 37 本。我们衷心感谢广大读者的支持和建议。

随着与 SAGE 出版社合作的进一步深化,我们又从丛书中精选了三十多个品种,译成中文,以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总序

2003年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课"社会科学里的统计学"(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课"社会科学中的定量分析"为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的"消费者"而非"生产者"。作为"消费者"，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复，但各有侧重。“社会科学里的统计学”从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂，与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程，他们当时大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在"格致方法 · 定量研究系列"丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海（夏季）和广州南沙（冬季）联合举办《应用社会科学研究方法研修班》，至今已经成功举办三届。研修课程设计体现"化整为零、循序渐进、中文教学、学以致用"的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余，历经近两年的时间，完成了三十多本新书的翻译任务，使得"格致方法・定量研究系列"丛书更加丰富和完善。他们是：东南大学社会学系副教授洪岩璧，香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多，硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林，原硕士研究生李兰，密歇根大学社会学系博士研究生王骁，纽约大学社会学系博士研究生温芳琪，牛津大学社会学系研究生周穆之，上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之，以及香港科技大学社会科学部硕士研究生陈佳莹，上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版，并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动，我在此致以诚挚的谢意。当然，每本书因本身内容和译者的行文风格有所差异，校对未免挂一漏万，术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议，以便再版时修订。

我们希望本丛书的持续出版，能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

第一次听到“虚拟变量”这个词的时候，许多定量研究方法的学生都会觉得有趣，但很快他们就会意识到，这个听上去“虚拟”的方法，在定量研究中却起着至关重要的作用。我们知道，在回归分析中，用定序或者名义变量作为自变量来进行回归分析，既不能有效地反映因变量与自变量之间的实际关系，而且又容易出现拟合不足的情况。然而，引入了“虚拟变量”的概念后，我们就可以在不违反测量相关假设的情况下，运用最小二乘法进行回归分析。

那到底什么是“虚拟变量”呢？简单地说，虚拟变量是由原先的定性变量构建出来的二分变量。对于二分法，通常需要 $G-1$ 个数字来涵盖所有信息，其中 G 为原先类别的个数。例如，在民意调查中，如果我们希望表达公民的政治兴趣（其中包括 3 个类别——非常同意、有点同意、不同意），研究者必须构建两个二分变量。假设它们分别为 X_1（编码 1 表示非常同意，0 表示除非常同意外的类别）和 X_2（编码 1 表示有点同意，0 表示除有点同意外的类别），如果 X_1、X_2 两个变量的编码都为 0，那么暗示了受访者所属类别为不同意。在这里，

"不同意"这个类别被设置成了底线，或者说是一个参照组，从而 X_1 和 X_2 的回归系数都是在其他组与该组比较后估计得到的。

但是为什么选择"不同意"作为参照组而不选其他类别，如"有点同意"呢？曾经使用过虚拟变量的研究者基本都遇到过这样的问题。在这里，Hardy 教授给出了明确的答案。在本书中，一个有关收入的、精心设计的例子贯穿全文，从一个简单模型（含有一个虚拟变量的回归模型，我们常常将其简化到均值差异的检验）到一系列复杂模型（含有多个虚拟变量、多个定量变量及多个交互项的回归模型）。所幸的是，通过严谨的语言叙述，这种复杂性可以用不同条件下所得的回归系数来表达。

对虚拟变量回归有了基本了解后，Hardy 教授还提出了有关虚拟变量回归的一些特殊问题。除此以外，她还对如何处理异方差性，在因变量取对数或者 logit 后，如何对回归系数进行诠释，如何在显著性检验下进行多重比较，如何进行效果编码和对比编码以及如何检验曲线性和如何进行分段线性回归作出了解释。

总之，本书以通俗易懂的语言，从不同角度对虚拟变量的用法进行了详述。在有关统计方法的书籍中，没有任何一个作者可以如此全面地诠释一个问题。可以说，这本书无疑是一部有关虚拟变量回归的重要著作。

迈克尔·S. 刘易斯-贝克

目 录

第1章

简 介

回归分析是定量分析中运用最灵活、最广泛的一种方法。一个典型的回归模型试图将因变量 Y_i 映射到一系列特定的自变量 X_i 上，并通过相应的线性函数来解释因变量 Y_i 的变异。利用最小二乘估计，我们可以得到一个预测方程，用来估计自变量的条件均值，即特定自变量组合下的 Y 的期望值，从而得到因变量的条件均值。当自变量像定量变量那样可测量时，我们可以假设其为一系列任意的相对零点且间隔大致相等的定量变量，此时，所有可能的 Y 的期望值都是无限的。此外，当因变量和自变量都是定量变量时，其相应的关系可用几何图形表示。

在二元回归中，我们预测 Y 为唯一自变量的函数，则两个变量之间的关系可由回归线直接表示。线上所有的点代表 Y 的条件均值。当有第二个自变量包含到函数中时，一维回归线扩展成二维，一个由南北方向和东西方向的线组成的平面生成了，此时代表 Y 的条件均值的是所有处于该平面上的点。由此可见，当自变量的数量增加时，这些原则是保持不变的，尽管其几何形态可能变得难以描述。

但如果所有用来预测的自变量都用间隔尺度来衡量，那么回归模型的有效性将会受到严重制约。我们研究的问题

经常涉及组差异,如社会学家感兴趣的对民族/种族差异、性别差异,或行为、态度及社会经济特征的区域差异的解释。又如,市场调研人员希望从人口统计数据中了解消费者偏好。研究人员常常想知道对于所有组别,自变量的影响是否一样,或者在同一关系的强度或方向上,组差异是否依然存在。由此可知,我们大多数的研究问题是为了区别各级因变量下的组差异以及不同自变量影响下的因变量的组差异。

当感兴趣的自变量为定性变量时(即"只在名义水平上测量"),我们需要一种方法,它既能定量地代表这种信息,又能防止将不切实际的测量假设强加于分类变量。例如,我们可以将职业分类按 1 到 12 进行编码(该分类用于人口普查中的单数代码),但我们不可以简单地说,职业的范围是从低值 1 到高值 12,因为这种描述是建立在假定的间隔相等的基本衡量标准上的。定义一系列虚拟变量可以使我们捕捉到分类方案里的分级信息,然后把此信息用到标准回归估计中。事实上,回归方程中的自变量可以是任意定性和定量预测因子的组合。

例如,"社会资源是通过收入进行分配的",这个现象既是那些对不平等感兴趣的学者所关注的焦点,也是那些努力为维持生活水平而奋斗的人民群众所关心的问题。我们关于社会公正的信念往往建立在对资源分布的认识上,以及是否有某些特定团体在分配过程中处于优势或劣势。我们知道,对于研究劳动收入分配中的歧视,有一种常见的方法,即首先确定一个组差异,比如男人和女人的差异或者黑人和白人的区别,以这个组差异作为在劣势群体的总效应,然后探讨加入其他决定性因素后,这个总差异如何变化,它是不是仍然维持不变? 通过此方法,那些形成于社会进程中的、可

察觉的不平等从而可被识别。

为了之后讨论统计方法时的连贯性，我会引用一个例子，即预测收入是个体特征的函数，并用定性或定量变量描述相应的个体特征。我所用的数据来自全国老年男性纵向调查。通过第一次入户结果[1]可知，在最初的样本中，我们的研究对象大约为美国1500万45岁至59岁且未收容到专门机构（如监狱、精神病院）的男性。在该例中，我们比较感兴趣的变量包括种族、职业（美国人口普查分类）、教育（受教育年限）和工作任期（在同一个雇主下的工作年限）。尽管其他变量，例如劳动力的供给、工作技能、健康等也可以被假设为（通过薪酬得到的）年收入的预测因子，但是对于此例，我们不予考虑，而用只含有四个预测因子的函数提供一个定性定量相结合的估测。通过讨论逐步复杂化的模型来阐述虚拟变量回归的方法，我会尽量解释清楚有关任意特定的虚拟变量的系数是如何随模型整体而变的问题。同时，我还希望通过这些努力，减少读者在不适用的情况下，对此方法进行演绎的可能性。

本书以讨论我们最初关注的问题——黑人和白人之间的收入差异（用"美元/年"衡量）开始。之后，我们会不断加入新的假设并逐步建立复杂的模型进行检验。我们所要估计的是，当控制了更多的自变量（包括定性的或定量的）后，黑人和白人之间的平均收入差异是否仍然存在。还有，各个自变量的净效应在黑人和白人中是否一样。最后，我们将使用虚拟变量回归的形式来估计种族对回归模型所有参数的具体影响。有关这个逐步深入的过程，我们将在第4章具体描述。尽管未必所有读者都对收入分配这个话题感兴趣，但是由于其中所涉及的方法比较直接简单，所以适合各个学科

背景的读者。此外,这些对模型的解释方式很容易扩展到其他实质性研究里,所以我还是选择了这个例子。本书第 5 章提供了一个有关对虚拟变量的替代性编码策略的简短描述。在第 6 章中,我们会把视线从单个问题上移开,而更加关注虚拟变量在其他研究里的运用。

第1节 | 多元线性回归回顾

随后的讨论均建立在有关单次方程回归模型的概念、偏相关的方法以及假设检验的基础上，这些读者都已比较熟悉。如果读者对这些还不太清楚，建议先阅读有关回归的介绍性书籍，其中有较早的卷本（比如，Berry & Feldman，1985；LewisBeck，1980；Schroeder，Sjoquist & Stephan，1986），还有一些基本的统计书籍（比如，Bohrnstedt & Knoke，1982；Cohen & Cohen，1983），这些书可为我们即将讨论的问题提供非常有用的信息。

在文中，我们还会用到一些符号语言，因此，现在来回顾一些基本的符号。假设我们有一个定量的因变量（Y_i），其为三个定量自变量 X_{1i}、X_{2i}、X_{3i} 的线性函数，则总的回归函数可写为：

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \\ &= \beta_0 + \sum \beta_k X_{ki} + u_i \end{aligned} \qquad [1.1]$$

其中，k 为第 k 个自变量，i 为第 i 个观测值。该方程表达了 Y_i 是 X_{1i}、X_{2i}、X_{3i} 以及随机误差项 u_i 的线性函数；β_0 是截距项，其可解释为当所有自变量均为 0 时，Y_i 的值；β_1 为总体偏回归系数，表示当控制了其他自变量后，X_{1i} 每变化一个单

元，Y_i 增加或者减少的量；β_2 和 β_3 同样也是总体偏回归系数，其分别是变量 X_{2i} 和 X_{3i} 的系数。由此可知，总体回归函数对于给定的自变量 X_{ki}，提供了 Y_i 的条件均值或期望值。因此，我们可以通过样本回归函数，用最小二乘估计来估测其所在的总体参数。

$$Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + B_3 X_{3i} + e_i \qquad [1.2]$$

每个回归系数 B_0、B_1、B_2、B_3 既是方程 1.1 中相对应的总体参数的点估计，也是统计量抽样分布的一个观测值。我们用 e_i^2 的观测值来估计总体方差和抽样分布里 B_0、B_1、B_2 和 B_3 的标准误，从而可以评估所得出的估测的显著性意义，进而对 Y_i 和 X_k 的关系作出结论。此外，标准差还可以用来构造区间估计，该区间通常被称为"置信区间"，其对评估有关假定的统计证据很有用。当如下假设都成立时，我们就可以用最小二乘法来分析这些样本数据了：

(1) $E(u_i \mid X_k) = 0$；即，在给定的 X_k 值下，u_i 的平均值为 0。

(2) $\mathrm{cov}(u_i, u_j) = 0$；即，对所有 $i \neq j$，干扰项之间是相互独立的。

(3) $\mathrm{var}(u_i) = \sigma^2$；即，对任意 X_k 的取值，u_i 的方差都是非负常数 σ^2，这也是同方差性的假设。

(4) $\mathrm{cov}(u_i, X_k) = 0$；即，干扰项和解释变量是相互独立的，彼此不相关。

在这些假设下，OLS 估计是最好的无偏估计，"最好"是因为

在所有线性无偏估计中,其方差最小。

异方差性的问题通常与截面数据(描述整体的单元群在一个特定时间点的数据)相关,自相关作用常常与时间序列数据(描述实体在一段时间内的数据)有关。而虚拟变量在研究截面数据和时间序列数据里,都扮演着非常重要的角色。在截面数据中,虚拟变量可以用来估计各群体之间的差异或者加入某一群体后,是否会改变其他解释变量的效应问题。同样,在时间序列数据中,虚拟变量可以用来确定两个时间段是否有区别,或检验不同时间段上,其解释变量所造成的影响的稳定性如何(Gujarati, 1970)。由于虚拟变量通常既可以在截面数据分析也可以在时间序列分析中定义观测组,因此,研究人员必须谨慎处理这两种情况下的异方差问题。在截面分析中,我们可以通过指定虚拟变量以获得组差异,但是,其前提是那些潜在的异质组的信息已经被合并了。如果这些群体的误差方差显著不同(即,如果我们违反了同方差性的假设),那么单个回归系数的显著性检验将会变得很不可靠。像这样类似的问题也可能在时间序列模型分析中出现,因为虚拟变量常用于检测两个或多个时间段的系数的稳定性,所以,如果误差方差在不同时间段中呈现出显著区别,那么异方差会使回归检验非常有争议(Maddala, 1992)。有关假设的讨论、违反假设的种种后果的详述以及有关处理这些违反假设的种种补救方法,在很多中级统计教材里都有提及。因这些均与虚拟变量的使用有关,所以我也会在本书中予以讨论,有关异方差性及自相关作用的问题,我们会在后文详述。

通过 R^2,即多重相关系数的平方,我们可以评估回归模

型对样本数据的整体拟合度。t 检验通常用来检验单个回归系数的统计显著性，为什么我们用 t 分布而不是 z 分布呢？其原因在于，我们一般不知道总体方差 σ^2 的值。因此，我们只能用样本的误差方差作为对总体方差的估计。当检验零假设，即其效应或局部效应是否等于 0 时，t 检验可以把参数估计的比率降低到其标准误。

由于三个自变量均包含在统计规范中，B_1、B_2、B_3 可用来估计 X_1、X_2、X_3 对 Y 的局部效应。一般来说，局部效应不等于当 Y 只对一个自变量回归时产生的二元效应，因为在一个给定的规范中，自变量通常相互关联或与 Y_i 共协方差。当其中一个自变量（例如，X_{1i}）与其他一个或多个自变量完全相关（即，该自变量是其他一个或多个自变量的线性函数）时，那么该样本估计是不确定的。直观地讲，我们可以把这种不确定性归因于缺乏"唯一的"信息："X_{1i}分布中的信息直接照搬方程的右边所包含的统计信息，当 X_{1i}没有提供任何净分布信息（例如，局部信息或唯一信息）时，我们不可能估计 X_{1i}对 Y 的净效应（不论局部或者唯一）。"这就是完美的多重共线性。从统计学上讲，当估测偏回归系数时，在我们可以明确"其他自变量被控制了"的意义后，该不确定性就可以被解释了。在统计学上，"自变量被控制"需要我们将与模型中其他自变量有关的变异从 Y_i 的分布中移除。由此可见，统计上的"控制"是一个分割变异的过程。在我们的样本回归函数中（方程 1.2），当确定 B_1 时，我们会移除 X_{2i}和 X_{3i}对 Y 的影响。换句话说，当 X_{2i}和 X_{3i}不能在样本中变化时，我们就看不到那些与 X_{2i}和 X_{3i}的变异相关的部分体现在 Y 或 X_{1i}的分布里。因此，在估计 X_{1i}对 Y 的偏效应时，我们不可

以把那部分变异，即与 X_{2i} 和 X_{3i} 的变异相关的部分变异考虑在内。从本质上来讲，X_{1i} 对 Y_i 的偏效应是基于两个残差分布——移除了 X_{2i} 和 X_{3i} 对 Y_i 的线性效应后的 Y_i 的残差分布，以及移除了 X_{2i} 和 X_{3i} 对 X_{1i} 的线性效应后的 X_{1i} 的残差分布。当存在完美的共线性时，X_{1i} 的残差分布是一个常数——0。

当我们把虚拟变量加入回归方程中后，回归估计的逻辑是不变的，即我们将根据因变量来预测条件均值，也就是说，通过把给定数值的自变量代入方程而得出 Y 的均值。其区别在于，虚拟变量的编码通常代表每个不同组群，或者根据是否有某个或某些特征，将虚拟变量分成有或者没有两种状态。因此，预测一个虚拟变量编码的特定组合的 Y 的期望值与预测组群的均值没有什么差别。该思想也同样适用于自变量是连续的时候。此时，“控制自变量”的过程对解释虚拟变量尤为重要。

第2章

构建虚拟变量

对分类数据编码要求构建完全穷尽且相互独立的类别。该原则同样适用于虚拟变量的构建。我们需要构造一个足够庞大的虚拟变量集合，从而将原先在定性范围内的所有信息都尽量表达出来。分类变量可以是两分或者多分的。一个有 j 个类别的分类变量需要 $j-1$ 个虚拟变量来获得初始差异集合中的所有分布信息。因虚拟变量通常为二分变量，所以我们通常用二进制编码(0，1)进行区分。所有在特定类别中的成员会被分配到代码 1 中，其他不在该类别的成员被分配到代码 0 中。根据这个编码原则，我们为一个给定的大类构造了一系列虚拟变量，原因在于在真实数据中，每个受访者只可对应大类中一个且只有一个编码为 1 的虚拟变量。我们可以把二进制编码想象成电气开关：编码 1 亮起时，表示一个给定类对一个受访者信息(例如，他/她是某一特定组群的成员或者他/她具有某一特定的特征)“开启”。对那些非成员来说，虚拟变量会切换到“关闭”状态(表示某些特征不存在)。

我们常用 $j-1$ 个虚拟变量来描述一个包含 j 个类别的定性变量，这样做的原因在于其直接符合古典线性回归的模型，自变量没有完全共线性的假设要求。也就是说，在一个

模型里,任何解释变量都不可能与其他解释变量存在完美共线关系。那么,假设以下例子中,我们用虚拟变量来表示种族。其中,被编码为1的虚拟变量(黑人)代表非洲裔美国人,若我们再加入第二个虚拟变量(白人)使其代表那些非非洲裔美国人,则在该模型中,就潜在地构造了两个自变量之间的完全线性关系,因为BLACK = 1 − WHITE。所以,白人这一虚拟变量对估计是多余且不必要的。

当原先的变量如之前提到的种族(黑人、白人)只有两类时,我们只需一个单一的虚拟变量便足以捕捉全部信息了。其中,没有用虚拟变量命名的类为参照组。如果原先的变量有两个以上的类,虚拟变量的数目就取决于在分析中需要比较的不同类的数目。以职业为例,在普查里,其通常以12位数编码来衡量。在该例中,我们以第12类作为参照组,那么,最多可以创建11个虚拟变量。为了解释得更清楚、更容易理解,我们将排除农场管理人员和雇农,并把剩余的组归为一类,此时,我们只需考虑六个类别:OCC_1(高级白领,如专家、经理等)、OCC_2(初级白领,如文员、推销员等)、OCC_3(技术工人,如木匠、水管工人、电工等)、OCC_4(操作工人,如焊工、织布工人及在生产制造中的装订工人)、OCC_5(非家庭服务工作者,如理发师、守卫、实习护士等)及OCC_6(劳工,如渔民、锯木工人、货车司机等)。根据前述的规则可知,这组含有六个详尽且互斥的类别需要五个虚拟变量来表达原先定性变量的所有信息,其中,五个虚拟变量分别表示不同的类,剩下的第六类,即被排除的类(未用虚拟变量命名所表示的类)为参照组。

第1节 | 选择参照组

在为一些分类变量编码时，我们必须选择参照组，即我们想把哪个或哪些组作为要比较的类。对每个分类变量，我们必须指定一个类作为参照组。在我们的例子中，如果选择“白人”作为参照组，那么虚拟变量“黑人”在二元回归里的系数就代表美国黑人相对于白人的平均收入是多少。换句话说，回归系数表达了黑人与白人之间的平均收入差异。那么如果以“白人”为参照组，则虚拟变量“黑人”在二元回归中的系数 B 可表示为如下：

$$B_{\text{BLACK}} = \overline{Y}_{\text{BLACK}} - \overline{Y}_{\text{WHITE}}$$

相反，如果非洲裔美国人为参照组，则白人为虚拟变量，此时，二元回归系数 B' 可表示为：

$$B'_{\text{WHITE}} = \overline{Y}_{\text{WHITE}} - \overline{Y}_{\text{BLACK}}$$

不论哪类被选为参照组，平均收入的差异绝对值是不变的。

为一个多分类变量选择参照组（例如职业）多少有些复杂。对于职业这一变量，其所有的虚拟变量的回归系数的估计都是相对于该参照组的。尽管我们可以使用任意一个组作为参照组进行回归估计，从而产生与其他类的比较，但是仍有些准则需要了解。以下这些准则对解释回归估计非常

有用。

第一,参照组应该明确界定。用其他剩余类可能不是一个好的选择,因为该选择无法确定“剩余”类的组成是什么。而且,我们感兴趣的组差异可能无法通过这些更相似的类与该剩余类的比较而反映出来。因此,选择一个明确的参照基准,可以将组差异清楚地表现在方程中,这点非常重要。

第二,当定性类别中存在隐含的次序关系时(例如职业),有些研究人员通常选择最低或者最高级别作为参照组,而其他人更倾向于选择中间的类。尽管前者可以提供一系列系数估计来解释那些相对最低或最高级别的类,但是后者却可以减少那些粗心的研究者光抓住一个统计上显著的系数(例如职业)而忽略了其作为一个多类别的预测因子,在整体上是否有着显著影响的可能性(该问题会在第4章结尾详细讨论)。

第三,一个参照组要包含足够多的事件及信息,从而可以更加合理、准确地估计组群的均值。有时,研究者为了使其他类的信息保持“单一”,可能会选择只含少量观察值的类作为参照组。但是我们要明确,这个战略只可以用在参照组是高密度分布的类上,即在该变量所有的类中,其观测数最多或者为最多之一。

读者需谨记在心的是,在统计层面上,参照组是可以任意选取的。假设参照组是研究者根据适当的解释和推理的过程选出来的,那么就不会有“错误”的选择。而在实际层面上,“最佳”的选择是最大限度地减少额外计算的数量,因为这些额外计算会产生最具实质性利益的信息。

表2.1提供了如上所述的编码过程。对于种族这个变

量，我们选择“白人”作为参照组；对于职业，我们选择“高级白领”作为参照组。受访者为非洲裔美国人的被编码为1，白人受访者被编码为0。OCC_2、OCC_3、OCC_4、OCC_5和OCC_6是五个虚拟变量，它们旨在捕捉变量职业的六类里的全部信息。其中，OCC_2代表初级白领，OCC_3代表技术工人，OCC_4代表操作工人，OCC_5代表服务业工作者，OCC_6为劳工。对于职业，事件3、事件14和事件15均为0，因为这些事件的受访者都是高级白领。而我们又可看出，事件3和事件14所有变量的编码在表中均为0，原因在于，这两个事件的受访者均是种族和职业这两个定性变量的参照组中的成员，即他们都是从事高级白领职业的白人。

表 2.1 种族及职业的虚拟变量的编码

事件	种族	职业	黑人	OCC_2	OCC_3	OCC_4	OCC_5	OCC_6
1	黑人	初级白领	1	1	0	0	0	0
2	白人	技术工人	0	0	1	0	0	0
3	白人	高级白领	0	0	0	0	0	0
4	黑人	操作工人	0	0	0	1	0	0
5	黑人	劳工	0	0	0	0	0	1
6	白人	初级白领	0	1	0	0	0	0
7	白人	技术工人	0	0	1	0	0	0
8	白人	服务业工作者	0	0	0	0	1	0
9	黑人	服务业工作者	1	0	0	0	1	0
10	白人	初级白领	0	1	0	0	0	0
11	白人	操作工人	0	0	0	1	0	0
12	白人	初级白领	0	1	0	0	0	0
13	黑人	技术工人	1	0	1	0	0	0
14	白人	高级白领	0	0	0	0	0	0
15	黑人	高级白领	1	0	0	0	0	0

所有隐含在种族和职业中的定性信息都可以被转化成可供计算的信息。通过转化，我们可以计算集中趋势、分散

程度、相关度及回归系数。我用 $j-1$ 个虚拟变量而不是原先含有 j 类的分类变量，其中一个重要的原因在于，每一个虚拟变量可以从原先的度量中提取一部分信息。例如，每个虚拟变量记录着一个职业特征的存在与否(例如，1 代表其劳工的职业特征存在，0 则代表其劳工的职业特征不存在)。我们没有从根本上改变包含在种族或者职业的信息中的内容，我们只是选择了一个可替换的形式来表述这种信息。因此，只要我们可以调整对回归系数的诠释，使它们与自变量所隐含的测量性质相一致，那么就可以说，我们的统计基础非常坚实而且牢固。

第 2 节 | 描述性统计

分布统计

由于虚拟变量通常与定性度量相关，那么那些关于各类别的频数及其所占比例的描述性信息就对描述变量分布非常有用。同样，集中趋势的两个最常用的度量——众数和平均值，也可以提供很多有用的信息。

虚拟变量均值可以告诉我们被编码为 1 的类占所有分类的比例。回想一下，这个比例其实是一个相对频数，它是通过给定的分类事件数除以所有事件数(n_i/N)得来的。让我们再回想一下那个计算均值的公式，即所有该度量下事件的数值的和除以事件的总数。假设所有的事件对一个虚拟变量只被编码成 0 或 1，那么加和所有事件的数值与统计所有被编码为 1 的事件数是等价的。因此，对于虚拟变量，比例公式与均值公式是等价的。

同样，对于包含连续度量的虚拟变量，其方差方程则与我们普遍运用的方差方程有关。

$$\left(\sum X_i^2\right)/N-\left(\sum X_i/N\right)^2=n_j/N-p_j^2$$
$$=p_j-p_j^2=p_j(1-p_j)$$

[2.1]

若 X_i 是连续的，方差公式为第一个方程等号左边的部分。当把同样的方程运用到虚拟变量中时，($\sum X_i^2$) 变为 n_j，即被编码为 1 的事件数。第二项均值的平方($\sum X_i/N)^2$ 变为被编码为 1 的事件的比例，如上所述，这两者也应该是等价的。因此，我们可以证明，虚拟变量的方差其实是被编码为 1 的事件的比例与被编码为 0 的事件的比例的乘积。

当事件均匀地在两类之间分布时，虚拟变量的变异最大。现在，让我们来讨论些有关公众舆论的问题。“你是否支持为公共教育增加税收?”对于这个问题，当人们的意见均匀分布的时候，对税收政策持反对意见的最多。这时，选中任意两个都为“支持”的概率是最小的。当观点趋同时，即同意或者反对的概率接近 100%时，意见的多样性(或者说变异)会随之下降。

相关性

研究者不仅对那些可以描述单变量分布的度量感兴趣，还对变量之间的相联度量感兴趣。尤其当要用列联表分析来调查定性变量之间的关系时，我们会通过检验离散变量分类的均值差异来估测定性和定量变量之间的关系。最初，我们会将其限制到三个度量里，因此，我们可以看到按职业和种族分类后的平均收入差异(见表 2.2)。

表 2.2 不同种族、职业下的收入平均值和标准差

	平均收入	黑人的百分比
种族		
白人 ($N=2290$)	7821.90 (4974.8)	
黑人 ($N=921$)	4619.00 (2428.1)	
职业		
高级白领(OCC_1) ($N=644$)	10702.10 (7166.5)	6.8
白人 ($N=602$)	10960.30 (7273.2)	
黑人 ($N=42$)	7001.80 (3874.5)	
初级白领(OCC_2) ($N=337$)	7680.90 (4228.7)	17.1
白人 ($N=279$)	8061.30 (4462.6)	
黑人 ($N=58$)	5850.80 (2039.9)	
技术工人(OCC_3) ($N=810$)	6945.00 (2864.9)	17.7
白人 ($N=665$)	7334.70 (2786.9)	
黑人 ($N=145$)	5157.80 (2526.0)	
操作工人(OCC_4) ($N=788$)	5553.90 (2454.1)	38.9
白人 ($N=481$)	6085.30 (2414.6)	
黑人 ($N=307$)	4721.40 (2281.5)	
服务业工作者(OCC_5) ($N=287$)	4434.4 (2352.0)	51.2
白人 ($N=139$)	4805.6 (2626.5)	

续表

	平均收入	黑人的百分比
黑人 $(N=148)$	4085.8 (2008.3)	
劳工(OCC_6) $(N=345)$	4090.0 (2020.1)	64.0
白人 $(N=124)$	4777.30 (1900.1)	
黑人 $(N=221)$	3704.30 (1986.6)	

通过对表 2.2 的描述性分析，我们会发现三个非常明显的趋势：第一，黑人的平均收入比白人低；第二，从高级白领到劳工，平均收入是逐渐减少的；第三，从高级白领到劳工，黑人的比例是逐渐增加的。现在，我们必须寻找合适的方法来总结这三个二元关系，检验它们的显著性，然后通过控制其他相关因素来估测这些关系。

我们已知虚拟变量的均值和方差均和 p_j 有关，那么，建立在样本方差和协方差上的相关度量也和虚拟变量的比例分布有关。大家知道，相关系数常用来测量变量之间相联程度，且建立在两个变量之间协方差上的度量又和两个变量分布的离散程度有关。因此，我们可以说，两个定性变量度量的相关性对原始分布中方差的数值很敏感，因为虚拟变量的方差是 p_j 的函数，且所涉及的虚拟变量相关性的强度会反映各类别出现频数的相对大小。

表 2.3 是对虚拟变量种族、职业及因变量收入的零阶相关性系数的估测，最右列列出了每个虚拟变量与收入的相关性。第一个数值(−0.313)表示黑人和收入的关系。负号说明虚拟变量编码为 1 的黑人与低收入相关，即美国非洲裔男

性的平均收入比其他男性的平均收入低。通过对相关系数加平方，我们可以算出收入的样本方差有多少是可以被种族解释的。从而我们知道，大约有10%的收入方差可以通过种族间平均收入差异所解释。

表 2.3　种族、职业和收入之间的相关性系数

	初级白领 (OCC_2)	技术工人 (OCC_3)	操作工人 (OCC_4)	服务业工作者 (OCC_5)	劳工 (OCC_6)	收入
黑人 OCC_2	−0.087***	−0.139***	0.131***	0.157***	0.272***	−0.313***
(初级白领) OCC_3		−0.199***	−0.196***	−0.108***	−0.119***	0.057***
(技术工人) OCC_4			−0.328***	−0.181***	−0.199***	0.007
(操作工人) OCC_5				−0.178***	−0.197***	−0.166***
(服务业工作者) OCC_6					−0.108***	−0.170***
(劳工)						−0.211***
平均值	0.106	0.250	0.244	0.089	0.107	6890
标准差	0.308	0.433	0.430	0.285	0.309	4622

注：*** 表示相关性系数在0.001显著性水平上统计显著。

由于在职业分类中需要多于一个的虚拟变量来捕捉职业的所有信息，因此我们用五个相关系数来描述职业与收入的关系，其中，每个都是针对一个特定的职业类别。这五个零阶相关系数的任何一个都可以估计该指定类（例如，OCC_5 中服务业工作者）同其他所有类合并的收入差异。例如，OCC_5 和收入之间的相关性系数为−0.170，它代表服务业工作者比非服务业工作者（$OCC_1+OCC_2+OCC_3+OCC_4+OCC_6$）的工资低。当我们把该系数平方后，就可以估计，有百分之多少的收入方差

是由于从事某一职业的男性比不从事该职业的男性挣得多或少这一事实引起的。这里，我们可知，有 2.89%的收入方差是由于服务业工作者比非服务业工作者的收入少引起的。还需注意的是，OCC_3(技术工人与非技术工人)和收入之间的相关性很小且不显著。从该度量可看出，技术工人与非技术工人的平均收入没有显著的不同，该发现与技术工人不论在职业还是收入分布上都处于中间水平的事实一致。

其余各列的相关性表示每两个虚拟变量之间的相关度。因为两个虚拟变量之间的相关度等价于 φ 系数，又 $\varphi=(\chi^2/N)^{1/2}$，所以它们也与 χ^2 有关。任意两个度量之间的关系都可以在一个 2×2 的表格里表示出来。我们看到表格第一行涵盖了黑人与任意职业类别的相关性信息。一方面，我们发现 OCC_2 与黑人呈现了负相关，这说明黑人在初级白领里的比例比在其他职业类别中的小；另一方面，黑人与 OCC_4、OCC_5 和 OCC_6 正相关，这说明黑人在操作工人中的比例(38.9%)、在服务性工作者中的比例(57.2%)和在劳工中的比例(64.0%)均比黑人不在操作工人中的比例(25.1%)、不在服务性工作者中的比例(26.3%)和不在劳工中的比例(24.3%)高。我们还可以从表格里看出，黑人与 OCC_6(劳工)的相关性最强，其原因在于黑人在劳工中的比例是最大的。[2]

偏相关

当其他自变量被控制时，我们可以通过偏相关系数估计一个因变量和一个自变量之间的关系。表 2.4 记录了当不断有自变量被控制时，OCC_2(初级白领)和收入(Y)之间的一

系列偏相关系数。

表 2.4 职业虚拟变量和收入的偏相关系数及半偏相关系数

$r_{y, occ2}$	$=0.057^{***}$		
$r_{y, occ2. occ3}$	$=0.060^{***}$		
$r_{y, occ2. occ3, occ4}$	$=0.011$		
$r_{y, occ2. occ3, occ4, occ5}$	$=-0.068^{***}$		
$r_{y, occ2. occ3, occ4, occ5, occ6}$	$=-0.171^{***}$		
	偏相关系数	半偏相关系数	半偏相关系数的平方
$r_{y, occ2. occ3, occ4, occ5, occ6}$	-0.171^{***}	-0.191^{***}	0.036
$r_{y, occ3. occ2, occ4, occ5, occ6}$	-0.271^{***}	-0.294^{***}	0.087
$r_{y, occ4. occ2, occ3, occ5, occ6}$	-0.369^{***}	-0.387^{***}	0.150
$r_{y, occ5. occ2, occ3, occ4, occ6}$	-0.337^{***}	-0.357^{***}	0.127
$r_{y, occ6. occ2, occ3, occ4, occ5}$	-0.378^{***}	-0.394^{***}	0.155

注：*** 表示相关性系数在 0.001 显著性水平上统计显著。

鉴于 OCC_2 和收入之间的零阶关系是根据初级白领与其他类的工作者的平均收入相比较而来的，那么一阶偏相关系数 $r_{y, occ2. occ3}$ 就是控制了技工这一类得来的。由于技工现在被控制，那么这个偏相关代表初级白领与除去初级白领和技工这两类工作者的平均收入差异的相关性。由于下一个系数控制了两个自变量（OCC_3 和 OCC_4，即技术工人和操作工人），因此是一个二阶偏相关系数。在该例中，偏相关系数并不显著，这说明了初级白领的平均收入与除去了其本身以及技术工人和操作工人后的工作者（高级白领、服务工作者和劳工）的平均收入没有显著差别，这个结果的产生无疑是由于高收入组与低收入组是通过它们的中距均值求得的。随着越来越多的职业类别被控制后，偏相关系数的阶数越来越高，数值为负的程度越大。最高阶数（或称“四阶偏相关”）控制了所有的职业虚拟变量，结果显示，初级白领的平均收入明显低于高

级白领(参照组)的平均工资,此时,高级白领是唯一未被控制的组。表2.4的下半部分为我们呈现了职业虚拟变量的所有四阶偏相关系数。每一行的偏相关系数都表示一对职业虚拟变量和收入之间关系,且该相关系数是通过消除其他变量的影响,比较高级白领(参照组)和指定职业类得来的。从上往下看,越往下,偏相关系数负的程度越来越大,这是因为当劳工和高级白领相比时,其收入差异的强度是最大的,而当初级白领与高级白领相比时,其收入差异是最小的。

表2.4的中间列和右边列记录了半偏系数和半偏系数的平方。我们知道半偏相关系数是建立在相关性和回归之间的有益桥梁。用于建立偏相关系数的、不断消除其他变量影响的剔除过程与用于建立偏回归系数的过程一样,会影响因变量和自变量的分布。然而,如果用半偏相关系数,本身对因变量有一定影响的自变量就不会因为其被控制而将这部分的影响剔除(Cohen & Cohen, 1983)。半偏系数的平方表示一个自变量对 Y_i 的可解释方差的唯一贡献。在这里,"唯一贡献"是指,Y_i 的方差只归因于一个自变量,而不与其他被控制的自变量分享。例如,表示收入与 OCC_2 的第一个半偏相关系数的平方数值0.036是在控制了 OCC_3、OCC_4、OCC_5 和 OCC_6 之后得出的。通过定义初级白领是与高级白领不同的职业类别后,我们可以解释3.6%的收入方差。换句话说,3.6%的收入方差可被解释是基于初级白领的平均工资比高级白领低这一事实。同样,15.5%的收入方差可以通过指定劳工与高级白领之间的收入差异来解释。因此,我们可知,保持其他条件一致,组间差异越大,所得到的可解释方差就越大。[3]

第3章

虚拟变量回归

在本章中，我们会探究四个含有虚拟变量的回归模型。最简单的模型可表述为因变量收入为一个虚拟变量的线性函数。第二个模型和第一个相似，表述了收入为一个解释变量/自变量的函数，不同的是，在该模型中，自变量是职业，而在第一个模型中，自变量是种族。然而，由于职业是多类别的，即其中含有两个以上的类，因此，我们需要构造五个虚拟变量。在第三个模型里，两个定性变量均会加入，因此我们可以讨论种族之间的收入差异是否可在与种族有关的职业差异里体现。在最后一个模型中，我们会进一步扩展自变量的数目，即定量的解释变量，连同职业和种族的虚拟变量都被包含在其中。

通过估计二元回归方程，我们可以观察当自变量数目从少到多时，因变量期望值的变化。从而体会，回归方程所代表的因变量期望值 $E(Y_i)$ 由一个单一点进一步拓展到一系列连续数值构建的一条线的过程。其中，线上的每一个点都估计了一个特定的 X_{ki} 条件下的 Y_i 的期望值，表示为 $E(Y_i | X_{ki})$。此时，期望值是连续的，原因在于，X_{ki} 本身是一个连续度量，其代表了无限的潜在数值。

当我们处理虚拟变量时，自变量为只有两种可能的数值

的离散度量。那么,在建模时,尤其是为含有一个虚拟变量的连续因变量构造函数时,我们无法作出回归线。因此,我们可对每个可能的数值计算出一个 Y_i 的期望值,即,当 $D_{ji} = 1$ 时,得到一个期望值;当 $D_{ji} = 0$ 时,再得到一个期望值。这些估测出来的数值相当于条件均值,即组群 j 的均值 Y_i。

请看下面三个模型:

模型 1: $Y_i = f(\text{种族}) = \beta_0 + \beta_1 \text{BLACK} + u_i$

模型 2: $Y_i = f(\text{职业}) = \beta_0 + \beta_1 \text{OCC}_2 + \beta_2 \text{OCC}_3 + \beta_3 \text{OCC}_4 + \beta_4 \text{OCC}_5 + \beta_5 \text{OCC}_6 + u_i$

模型 3: $Y_i = f(\text{种族,职业}) = \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 + \beta_4 \text{OCC}_4 + \beta_5 \text{OCC}_5 + \beta_6 \text{OCC}_6 + u_i$

第 1 节 对含有一个虚拟变量的模型进行线性回归

在模型 1 里，通过收入对虚拟变量黑人进行回归，确定种族是否为收入的一个重要预测因子。表 3.1 列出了其回归结果。对这些连续和离散的自变量回归系数的合理解释要看常数项(B_0)，其代表当所有的自变量都为 0 时，Y_i 的期望值；同时，还要看 B_1，它代表 X_k 每变化一个单元，Y_i 期望值的变化。当 X_k 是连续的时候，Y_i 的分布也是连续的，因此，其回归系数也可表示为斜率。相反，当 X_k 为虚拟变量时，Y_i 的变化是随每次 B_k 单位的变化而变化，与之前不同的是，该变化基于是否为指定类别的成员的相关定义，因为虚拟变量一个单位的变化(从 0 到 1 或者从 1 到 0)反映了它是否属于某个指定类别。

在此例中，虚拟变量黑人的回归系数是负值。这说明，黑人的预测收入比白人的预测收入少 3202.90 美元。Y_i 的预测收入是通过简单的加减计算出来的。当 BLACK = 1 时，预测收入等于 $B_0 + B_1$，或者说 7821.9 − 3202.9 = 4619 美元；当 BLACK = 0 时，其预测收入就等于 B_0 或者 7821.9 美元。读者可以发现这些预测值与表 2.2 中所列出的组群均值相等。

表 3.1 模型 1、模型 2、模型 3 的回归结果

	模型 1	模型 2	模型 3
常数	7821.9 (91.9)	10702.1 (160.8)	10811.4 (158.9)
黑人	−3202.9 (171.6)		−1676.0 (172.4)
OCC_2		−3021.2 (274.4)	−2842.1 (271.1)
OCC_3		−3757.1 (215.5)	−3566.4 (213.3)
OCC_4		−5148.2 (216.8)	−4604.5 (220.9)
OCC_5		−6267.7 (289.7)	−5512.7 (295.9)
OCC_6		−6612.1 (272.3)	−5647.8 (286.2)
R^2	0.09792	0.22400	0.24624
F	348.3	185.0	174.1
R^2 的增量($R_3^2-R_1^2$)			0.148
变化过的 F			126.1

模型 2 回归系数的方差、协方差矩阵

	OCC_2	OCC_3	OCC_4	OCC_5	OCC_6
OCC_2	75309.07				
OCC_3	25870.70	46439.50			
OCC_4	25870.70	25870.70	47013.76		
OCC_5	25870.70	25870.70	25870.70	83922.03	
OCC_6	25870.70	25870.70	25870.70	25870.70	74162.67

注:括号里为所估计的回归系数的标准误。

虚拟变量的显著性检验要遵循标准化过程。黑人的回归系数测量了黑人的期望收入相对于白人如何。因此,黑人回归系数的标准误提供了白人与黑人的期望收入差异的标准误。当检验零效应即由组群差异引起的期望收入差异不

存在这一零假设时，t 检验为回归系数与标准误的比率。同样，由于模型 1 包含了一个单一自变量，那么 F 检验在此也是对零假设的检验，其值为 t 值的平方。R^2 说明，种族差异解释了大约 10%的收入中的方差，这一点在之前的零阶相关系数检验中就已得知。

该例描述了当自变量为虚拟变量和自变量为定量变量时，其回归结果解释的相似与不同。常数项估计了参照组（白人）的期望收入；B_1 估计了虚拟变量的特征对期望值的影响（例如，黑人对其期望收入的影响），该影响捕捉了黑人与白人之间的收入差异。因此，零假设（$\beta_1 = 0$）可表述为：$H_0: \mu_{BLACKS} - \mu_{WHITES} = 0$ 且对该模型的 B_1 的 t 检验和对模型 1 的 F 检验是相同的。[4]

第2节 | 对含有多个虚拟变量的模型进行回归

模型2估计了收入作为职业的函数，其中职业由五个虚拟变量表示。回归结果显示在表3.1的中间列。与模型1的解释一样，常数项10702.1为高级白领(参照组)的期望收入。其他的回归系数估计了从事相应职业类相对于高级白领的差异程度，从OCC_2的回归系数可以看出，初级白领收入平均比高级白领少3021.20美元，为7680.90美元。相比之下，劳工平均比高级白领少挣6612.10美元，只有4090美元。

用一系列虚拟变量比用单一虚拟变量更能捕捉各职业组之间的差别信息，那么相应的职业对收入影响的显著性检验应该为该模型的F检验。模型2的零假设可以写成$\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=0$，$F$检验是检验所有职业的期望收入是否都是一样的。另外，因为F检验可以表达为R^2与k个自由度的商和$1-R^2$与$N-k-1$个自由度的商的比率，其中，k等于自变量的数目，因此，F检验还可以看成是对R^2的显著性检验。因此，拒绝零假设说明一个非零的收入方差可被受访者的职业所解释。基于模型2的回归结果，我们有：

$$F_{53205}=\frac{0.22400/5}{(1-0.22400)/3205}=185.0$$

该数值在 0.001 的显著水平下非常显著。[5]建立了职业的统计显著性之后,我们现在可以转移到对单个回归系数的 t 检验上,从该检验我们可以看到,每个职业类的期望收入都与参照组呈现出显著不同。

第3节 | 估计类别之间的差异

t 检验同虚拟变量的回归系数一起，使我们可以检验相对于参照组，某职业类别带来的影响的显著性如何。然而，我们无法立即知道所比较的类别之间是否不同。例如，OCC_6（劳工）的回归系数负的程度最大，因此其期望收入也就最低。但是，我们如何知道劳工的期望收入确实是比服务业工作者或者操作工人低呢？

因为 $\beta_j = E(Y_i \mid OCC_j = 1) - E(Y_i \mid ref)$，所以期望收入在某两类之间的差异等于它们的回归系数之差（$\beta_j - \beta_k$），其中，β_j 代表第 j 类的虚拟变量的回归系数，同样，β_k 代表第 k 类的虚拟变量的回归系数。为了检验 OCC_4 和 OCC_6 的差异，即比较作为操作工人和劳工的差异，或者说，相对于劳工，操作工人所带来的影响，我们就必须用一个 t 检验来估计回归系数之间的差异：

$$t = (B_j - B_k)/[\mathrm{var}(B_j) + \mathrm{var}(B_k) - 2\mathrm{cov}(B_j B_k)]^{1/2} \quad [3.1]$$

因为回归系数方差正好是标准误的平方，所以它们本身就容易得到。另外，许多统计软件包都有计算回归系数的方差、协方差矩阵的选项，研究者可以更加轻松灵活地完成这些额

外检验。[6]

将 OCC_4 和 OCC_6 的估计值代入方程 3.1，我们有：

$$
\begin{aligned}
t &= -6612.1-(-5148.2)/\\
&\quad [74162.7+47013.8-2(25870.7)]^{1/2}\\
&= -1463.9/263.5 = -5.56.
\end{aligned}
\quad [3.2]
$$

在常用显著性水平 $\alpha=0.05$ 下，t 的临界值为 ± 1.96，从而我们可以推断劳工带来的影响确实与操作工人带来的影响不同，也就是说，劳工和操作工人确实有不同级别的平均收入。

第4节｜第二个定性度量的加入

当我们回头看表2.2中组群的均值时，可以发现，从高级白领到劳工，下降的不仅仅是平均收入，同时还有白人的比例。我们想知道，当控制了职业的收入差异时，种族差异是不是还会在收入中存在。要回答这个问题，就需要检验虚拟变量黑人的偏回归系数。表3.1最右列的模型3给了我们一个比较满意的答案。从常数10811.4可以得出当所有自变量为0时的期望收入，或者说，该常数即白人高级白领的期望收入。黑人的回归系数－1676.0表示，在我们考虑了同职业类别有关的收入方差和黑人在各职业类别中不是均匀分布的事实后，黑人的平均收入比白人少1676美元。尽管这个值比模型1中估计的小一些，但是该估计的收入差异在0.001显著性水平下非常显著。黑人的回归系数强度随职业由高到低逐渐下降，该现象说明黑人的平均收入比白人低的一个原因在于，黑人大体上都集中在低收入职业类别里。同样，如果控制了收入和职业分布中的种族差异，偏回归系数连同职业虚拟变量一起，可以估计出每个指定类别的成员对期望收入的影响大小。

为了检验控制了职业后的种族局部效应或者控制了种族后的职业局部效应是否仍在统计上显著，我们还要进行F

检验。另外，与其单靠 F 检验来检验整个方程，不如在控制了其他变量后，用增量 F 检验来检验一个或一组分类变量的解释功效。例如，我们可以把模型 3 看成模型 1 和模型 2 的结合，我们在模型 3 里加入了代表职业类别中种族差异的虚拟变量。和之前一样，职业定位的解释功效是由一组虚拟变量而不是一个虚拟变量来捕捉的，因此，我们可以通过比较模型 1 和模型 3 的 R^2 值，或者平均回归平方和来估计职业的分布。零假设在这里可表示为 $\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=0$，换句话说，一旦我们控制了收入和各职业类别中的种族差异，期望收入在所有的职业类别中都是相等的。该 F 检验的公式可表述为：

$$F=\frac{(R_3^2-R_1^2)/(k_3-k_1)}{(1-R_3^2)/(N-k_3-1)} \qquad [3.3]$$

其中，R_3^2 是模型 3 的 R^2 值，R_1^2 为模型 1 的 R^2 值，N 为事件数目，k_1 和 k_3 分别为模型 1 和模型 3 里自变量的数目。分子部分表明了相对于模型 1 和模型 3 自变量的数目差异，由职业类别影响所带来的 R^2 增量。分母部分为当种族和职业都包括在内后，所剩的、不能被解释的方差的比例与相应的自由度的商是多少。在该例中，加上从 OCC_2 到 OCC_6 观测所得的 R^2 增量为 0.14832，然后我们还需算出其除以五个自由度后所得的值。因此，在控制了种族后，用 F 检验算出职业类别的显著性为：

$$F=(0.14832/5)/(0.75376/3204)=126.1$$

第5节 | 期望值

由于种族包含两类,职业包含六类,那么他们一起可以生成共12个不同的组群。模型3针对这12组群估计了他们的期望收入。这些估计值和12个组群的均值相等,即与各个种族在每个职业类别中的期望收入均值相等。通过用表3.2中描述的回归系数集合,读者可以自己计算出各组群的收入期望值,然后和表2.2中列出的数值进行比较。表2.2中列出各组群的期望值是根据模型1估计出的种族参数,或者是模型2估计出的职业类别参数所得出的。和之前的结果不同的是,从模型3算出来的12个组群的期望收入和表2.2中的数值非常不匹配,这是为什么呢?

这是由于在模型1和模型2中,我们把所有检验限制在一维中,即种族或者职业。当将观测拓展到模型3时,我们其实是基于一个简单假设,即"黑人对各职业类别的影响是一样的(例如,黑人和白人之间的期望收入有差异)"和"职业间的收入差异对黑人和白人也是一样的"。我们知道,当计算职业类别中黑人和白人的期望收入有差异时,这个假设已经开始奏效,即不论职业,黑人和白人工作者的差别总是1676(B_1 的值);不论种族,服务业工作者和高级白领之间的差异总是-5512.7(B_5 的值)。这种等价性影响是模型阐述

的结果。

表 3.2 模型 3 的收入预测值

	Blacks	Whites
OCC_1	$B_0+B_1=9135.4$	$B_0=10811.4$
OCC_2	$B_0+B_1+B_2=6293.3$	$B_0+B_2=7969.3$
OCC_3	$B_0+B_1+B_3=5569.0$	$B_0+B_3=7245.0$
OCC_4	$B_0+B_1+B_4=4530.9$	$B_0+B_4=6206.9$
OCC_5	$B_0+B_1+B_5=3622.7$	$B_0+B_5=5298.7$
OCC_6	$B_0+B_1+B_6=3487.6$	$B_0+B_6=5163.6$

当把简化过的假设加入到模型 3 的规范中后，经验上是否行得通呢？事实上，通过了解各组群的期望收入与表 2.2 中组群均值差异，我们可能想改变这个设定。接下来在模型 4 中，我们会提供一个比较正式的检验，其结果会告诉我们哪个模型更好，是模型 3 还是允许差异效应(例如，交互作用)的模型。

第6节｜在模型设定中加入定量变量

用回归分析的好处是，即便一些自变量是分类变量，建模过程还是非常灵活的。基于这点，我们会将最初的模型限制在虚拟变量回归因子中，这样读者就可以习惯虚拟变量回归系数的解释方法了。在该部分，我们准备把定量和虚拟变量回归因子都加入观测中。因此，我们提出了模型4——期望收入是种族、职业、教育和工作任期的函数。

$$
\begin{aligned}
\text{模型 4: } Y_i &= f(\text{种族，职业，教育，工作任期}) \\
&= \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 \\
&\quad + \beta_4 \text{OCC}_4 + \beta_5 \text{OCC}_5 + \beta_6 \text{OCC}_6 \\
&\quad + \beta_7 \text{EDUC} + \beta_8 \text{TENURE} + u_i
\end{aligned}
$$

EDUC和TENURE都是用年来测量的定量变量。由表3.3可见模型4的回归结果。

我们可以看出，该模型估计结果的常数比之前的估计都小。更重要的是，模型设定的变化也改变了其实质性的意义。现在的常数估计的是那些没受过教育且工作任期为0的白人高级白领的期望收入，这些特征几乎是不可能存在的。虚拟变量黑人的回归系数现在表示当把职业、教育和工作任期这些对收入方差有影响的变量剔除后，黑人和白人之

间的期望收入差异(1188.10美元)。虚拟变量职业的回归系数估计了在控制了其他自变量之后,每个职业类别相对于参照组的期望收入的净差异。比如,初级白领平均比高级白领的收入少2316.10美元等等。同样,当保持种族、职业和教育不变时,每增加一年的工作任期,收入可增加84.70美元。对于教育,每增加一年的教育经历,期望收入会增加282美元。

由于另外两个定量变量的加入,我们可以把模型4想象为,其生成了一系列的回归平面,并引入了截距、斜率和偏斜率的概念。在文中,我们还可以将虚拟变量的回归系数表示成不同的截距。由于斜率或者偏斜率只存在于定量自变量中,因此,与EDUC和TENURE有关的回归系数提供了对偏斜率的估计。图3.1描述了模型4的回归结果。为了便于读者比较各组群之间教育和工作任期的截距、斜率和偏斜率,图3.1没有用三维空间图,而是用二维直线图对其进行了比较。又由于教育和工作任期均用年来衡量,则我们可以用同一度量对它们的局部效应进行比较。

对于所有受访者,由于模型设定把估计值限制到教育年限和工作任期所带来的平均影响里,因此所有实线均有着相同的斜率(282美元/年),所有的虚线也有着相同的斜率(84.70美元/年)。然而,Y截距是可以因组群的不同而不同的,因此在图中有12个不同的截距。由表3.2可见,计算这些截距时,我们用了相同的回归系数组合。但是,因为回归系数本身随着模型设定拓展而改变,那么,计算出来的截距自然会与表3.2中列出的预测收入有所不同。从每个"按种族特征分类的职业组"(共12个)的截距中分出一个实线和一

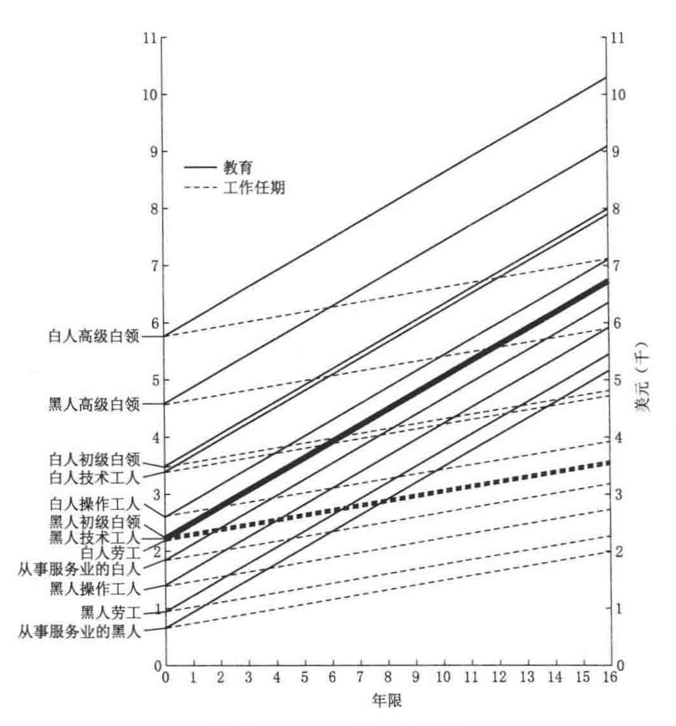


图 3.1 模型 4 的回归结果

个虚线，其分别代表教育年限和工作任期的偏效应。那么，对每个"按种族特征分类的职业类别"，我们就可以辨别出：(1)合适的 Y 截距，即有 0 年教育经历和 0 年工作任期的给定组群成员的 Y 的期望值；(2)对于一个特定的组群，每增加一年的教育经历，Y 净增加的期望值（实线）；(3)对于一个特定的组群，每增加一年的工作任期，Y 净增加的期望值（虚线）。就如模型 3，"各组群的等价性"仅是嵌入在模型设定中的一个假设，其经验性仍有待检验。

表 3.3 模型 4 的回归结果

常数	5761.1 (359.0)	OCC_6	−3606.8 (306.4)
黑人	−1188.1 (169.4)	教育	282.0 (23.1)
OCC_2	−2316.1 (261.8)	工作任期	84.7 (6.6)
OCC_3	−2343.7 (223.7)	R^2	0.31459
OCC_4	−3166.6 (237.5)	F	183.7
OCC_5	−3918.5 (299.9)	R^2 的增量(相对于模型 3) 变化过的 F	0.068 159.7***

注:括号里为标准误。
*** 表示相关性系数在 0.001 显著性水平上统计显著。

第4章

估计组影响差异

上一章的模型是拓展了模型设定中的自变量数目和种类后得到的。我们知道，所有含多变量的模型都有一个简化的假设，即任意自变量直接对因变量的影响与该自变量通过其他自变量对因变量的影响是一样的。换句话说，我们没有包含任何交互项来检验职业、教育或者工作任期是否在黑人和白人之间有所不同。在本章中，我们将建立一个新模型，通过引入交互项来检验假设的有效性，然后来回答之前讨论的两个问题——违反回归模型假设的后果和用非独立检验做多重比较的替代方法。

估计组群间的平均影响可以提供一个有用又简单的关系描述。然而有时，一个自变量(X_i)通过第二个自变量(Z_i)的分类或数值所产生的对因变量(Y_i)的影响是不同的。当 X_i 和 Y_i 之间的关系由变量 Z_i 决定时，我们就需要调整模型的设定，使 X_i 和 Y_i 之间的关系相对于 Z_i 而改变。检验这种差异效应需要用交互项，即包括在模型设定里的两个或多个自变量的乘积。

交互项可以定义为两个定量变量的乘积，也可以是两个虚拟变量的乘积，或者是一个定量变量和一个虚拟变量的乘积。另外，更复杂的交互项可以包括两个以上的变量。根据文献，早期对该系列研究有卓越贡献的包括 Jaccard、Turrisi

和 Wan(1990)，当所有变量都是连续度量时，他们为所出现的交互作用提供了非常完美的解释。但是在本章，我们关注的是其他的交互组合。

可以考虑构造一个包含两个虚拟变量(D_{1i}、D_{2i})的交互项来衡量两个二分的定性变量，比如性别和婚姻状况。我们可能会假设，作为女性所带来的影响可能会由其是否结婚而决定。在此情况下，我们会检验作为女性所带来的影响大小、已婚所带来的影响大小和交互项，包含变量 D_{1i}、D_{2i} 的乘积，即 $D_{1i} \times D_{2i}$。这样，当一个受访者为已婚女性时，其交互项就等于1，交互项的回归系数估计了已婚女性和其他受访者的不同影响程度。其中，其他受访者包括了已婚男性、未婚女性和未婚男性。

现在，我们可以考虑另一个交互项，它由一个测量年龄的定量变量 X_i 和一个性别虚拟变量 D_{1i} 所定义。在这里，我们可能会假设年龄所带来的影响由性别决定。因此，我们会试图用模型来检验年龄所带来的影响(X_i)、性别所带来的影响(D_{1i})以及交互项($X_i \times D_{1i}$)的影响。在这里，交互项对所有男性都为0；而对女性，则会默认为她们的年龄。该交互项的回归系数估计了年龄对女性的影响小于(或大于)年龄对男性的影响。

对于检验有关老年工作者收入数据的情况，图3.1可以帮我们把几种差异效应概念化。从图中我们可以看出，由于模型3和模型4的设定不同，使得黑人和白人之间的收入差异对所有职业类别都是一样的。那么，黑人的收入劣势会不会在高级白领中比较大，而在劳工中比较小呢？换句话说，Y 截距之间的差异在高级白领中是不是比在劳工中更大呢？这个问题的一个更普遍的问法是："种族的影响是不是在各

职业类别中不同？或者说，是不是从事不同职业对黑人和白人的影响是不同的？”

为了对那些包含定性变量的交互项作出更好的解释，我们可以先来看看表 4.1，其展示了收入、种族和职业之间的三种可能关系。在三个部分里，均列出了不同种族在各职业类别里的平均收入，而且，每一列的边际值还列出了该职业类别的平均收入。我们发现，表格的前两个部分用的是可以说明交互作用类型的假设数据，且相同职业类别的均值是相同的。知道了这些，我们可以进一步发现表格最右端的黑人和白人之间的收入边际差异不总能准确地捕捉到每部分内种族、职业和收入之间的关系。

表 4.1 交互项可能的种类

	高级白领	初级白领	技术工人	操作工人	服务业工作者	劳 工	
无交互作用[a]							
黑人	7708	5029	4315	3599	2883	2939	4412
白人	10911	8232	7518	6802	6086	6142	7615
	10702	7681	6945	5554	4434	4090	
包含强度和方向差异的交互作用[a]							
黑人	7002	5851	5158	6111	5120	4628	5470
白人	10960	8061	7335	5198	3704	3130	7479
	10702	7681	6945	5554	4434	4090	
包含强度差异的交互作用[b]							
黑人	7002	5851	5158	4721	4086	3704	4619
白人	10960	8061	7335	6085	4806	4777	7822
	10702	7681	6945	5551	4434	4090	

注：a. 构造的数据。
b. 真实的数据。

表格第一部分罗列的数据是没有交互作用的。当交互作用不存在时，处于边际的黑人和白人的平均收入差异为

3203美元，其与控制了职业类别后，黑人和白人在每一列内的平均收入差异的大小是一样的。读者可以通过用第二行每列的均值扣除第一行的均值来证实这点。在该情况下，所估计的种族在一个回归模型里的影响，即对每个职业组的平均影响(比如模型3)，就可以提供一个比较准确的黑人、白人间的收入差异，因为种族对所有职业类别的影响是一致的。

表4.1的第二部分仍然是基于构造的数据，说明了种族影响在强度和方向上的不同。该交互类型表明了，种族的影响不仅在不同职业类别中不同，即黑人和白人之间的平均收入差异强度不同，而且其影响方向也不同。假设样本中组群均值就如表格4.1中所列的一样，种族的收入边际差异或者中距的平均影响为2009美元，从而可知，白人平均收入稍高。但是当我们再仔细观察表格的该部分，以指定职业组间的比较差异为条件，黑人和白人之间的收入差异的强度是不同的。这个差异在高级白领中最大(3.958美元)，在操作工人中最小(913美元)。由于在各职业类别中，黑人和白人之间收入差异并不一致，因此我们需要指定一个交互项。同时，可以发现该差异不仅在强度上不同，其方向或者符号的正负也不总是相同。只有在高级白领、初级白领和技术工人中，白人的平均收入才比较高；相反，在操作工人、服务业工作者和劳工中，黑人的平均工资相对比较高。因此该交互作用标识了收入差异在方向上的逆转，暗示了黑人和白人收入的边际差异隐藏了种族差异的重要部分。如果将各职业类别的影响平均化，那么在初级白领和技术工人中，其种族影响是合理且准确的。但是对于高级白领，平均化低估了种族差异，而且对于其他职业类别，平均化还错误地估计了差异的方向。

通过表 4.1 的第三部分,我们可以比较实际数据和构造数据的分布。这些收入均值也可以在表 2.2 中找到。在实际数据中,黑人和白人之间的收入边际差异为 3203 美元,我们已经在第一个回归模型里估计出该值。但是当我们把视线放到每个职业内时,会发现黑人和白人之间的收入差异是由职业类别决定的,该差异在高级白领中最大(3958 美元),而在服务业工作者中最小(720 美元)。然而,收入差异总是在一个方向,即黑人工作者的平均收入总是比较小。在处理这种交互项时,其平均差异影响的方向通常是正确的,只是在一些职业类别中比较小,而在其他类别中比较大。

第1节 | 解释交互效应

为了检验交互效应，我们需要一个可以估计差异效应和确定其显著性的模型设定。要达到这个目标，我们就需要构造五个乘积项，然后把这五个项加入模型设定中。通过模型5，我们可以检验不同种族在不同职业中的差异效应，或者说不同职业在不同种族中的差异效应。

$$
\begin{aligned}
\text{模型 5：} Y_i &= f(\text{种族，职业，教育，工作任期}) \\
&= \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 \\
&\quad + \beta_4 \text{OCC}_4 + \beta_5 \text{OCC}_5 + \beta_6 \text{OCC}_6 + \beta_7 \text{EDUC} \\
&\quad + \beta_8 \text{TENURE} + \beta_9 \text{BLOCC}_2 + \beta_{10} \text{BLOCC}_3 \\
&\quad + \beta_{11} \text{BLOCC}_4 + \beta_{12} \text{BLOCC}_5 + \beta_{13} \text{BLOCC}_6 + u_i
\end{aligned}
$$

新变量 BLOCC_2 到BLOCC_6 是由虚拟变量黑人与每个职业虚拟变量相乘而得来的。如果受访者既是黑人又属于初级白领，那么BLOCC_2 被编码为1。因此，由BLOCC_2 的回归系数估计出的平均收入的增加或者降低，只能应用到黑人初级白领这个组群。

模型5的结果在表4.2里已列出。最初，我们可能想知道是否要在模型拟合中允许种族和职业的差异效应而使统计显著性提高。我们可以就这个问题，通过方程3.3的 R^2

增量检验，比较模型 5 与模型 4 的结果可知：

$$F_{5\,3197} = \frac{0.00679/5}{0.67862/3197} = 6.4$$

表 4.2 模型 5 的回归结果

常数	5794.8 (358.7)	TENURE	84.0 (6.6)
BLACK	−3793.3 (610.1)	$BLOCC_2$	1501.2 (823.0)
OCC_2	−2274.9 (280.2)	$BLOCC_3$	2326.2 (705.0)
OCC_3	−2418.4 (232.7)	$BLOCC_4$	2984.8 (672.5)
OCC_4	−3427.2 (256.3)	$BLOCC_5$	3528.0 (761.0)
OCC_5	−4513.4 (372.5)	$BLOCC_6$	3383.9 (747.3)
OCC_6	−4202.8 (399.0)	R^2 F	0.32138 116.46
EDUC	292.9 (23.1)	R^2 的增量($R_3^2 - R_1^2$) 变化过的 F	0.007 6.42***

注：括号里为回归结果的标准误。
*** 表示回归系数在 0.001 显著水平上显著。

该 F 值在 0.001 显著水平上统计显著。尽管所增加的解释功效没有显著到可以拒绝的地步，但是 F 检验的确告诉我们，样本量大，估计的差异效应更加合理准确。

现在我们再看对回归系数的解释。模型 5 的常数项与模型 4 中的一样，估计了有 0 年教育经历和工作任期的白人高级白领的预测收入。另外，EDUC 的回归系数告诉我们，在控制了工作任期、种族、职业以及职业内不同的种族影响后，教育对收入的平均影响是多少。TENURE 回归系数的解释与之类似，在此就不详细说明了。

黑人和职业虚拟变量的回归系数看上去像模型 4 的延续，其实不然。由于引入了种族×职业的乘积项，它们的意义出现了变化。当我们继续专注于由不同种族在不同职业中形成的 12 个组群时，可以通过将各回归系数映射到其所在的组群中，以弄清楚每个特定的回归系数所扮演的角色是什么。该映射结果已在表 4.3 中列出，根据组群中的个体在虚拟变量中(包括交互项里)编码为 1 的情况把回归系数加入到特定组群的估计中。为简化起见，我们还是将 EDUC 和 TENURE 的回归系数设定为 0。

表 4.3 不同种族预测收入系数

	白 人	黑 人
高级白领	B_0	B_0+B_1
初级白领	B_0+B_2	$B_0+B_1+B_2+B_9$
技术工人	B_0+B_3	$B_0+B_1+B_3+B_{10}$
操作工人	B_0+B_4	$B_0+B_1+B_4+B_{11}$
服务业工作者	B_0+B_5	$B_0+B_1+B_5+B_{12}$
劳工	B_0+B_6	$B_0+B_1+B_6+B_{13}$

从黑人的回归系数开始看，B_1 估计了黑人高级白领和白人高级白领之间的期望收入差异。与模型 4 不同的是，其提供的不再是黑人在所有职业类别中的平均影响。系数的 t 检验是对零假设的检验，即在控制了由教育、工作任期带来的方差后，黑人高级白领与白人高级白领的期望收入是一样的零假设。换句话说，即在高级白领中，黑人对期望收入并没有显著影响。由于该回归系数的 t 值为-6.22，因此零假设可以被拒绝。很明显，在高级白领中，当控制了模型中的其他因素后，黑人平均收入显著地低于白人。

同样，职业虚拟变量的回归系数也不再表示黑人和白人

在某一特定职业类别相对于参照组的平均影响。我们用 B_2，即 OCC_2 的回归系数来估计白人初级白领与白人高级白领之间的期望收入差异，从结果中，我们知道，白人初级白领平均比白人高级白领的收入低 2274.90 美元。同样，白人操作工人平均比白人高级白领的收入低 3427.20 美元。换句话说，一旦指定了乘积项，原来变量的系数（如黑人，还有 OCC_2 到 OCC_6 的系数）成为包括了参照组影响的比较。比如，B_1 测量的是黑人高级白领的影响；B_2 到 B_6 测量的是白人在除了高级白领的某一特定职业类别中的影响。t 检验连同 OCC_2 到 OCC_6 的回归系数一起，测量的是白人工作者在各职业类别的收入差异显著性。表 4.2 的结果表明，对于白人，高级白领与其他职业类别的估计差异显著不同。

乘积变量的回归系数估计了不同种族从事不同职业的差异效应。同样，我们还可以通过这些回归系数来估计黑人在各职业类别里的差异效应。为什么这两种说法都可以接受呢？通过观察表 4.3，我们可以给出问题的答案。白人受访者从事初级白领和高级白领工作的预测收入差异可通过 B_2（-2274.9）来表现，对于黑人，这个差异可通过 $B_2 + B_9$（$-2274.9 + 1501.2$）来表现。因此，B_9 估计的是相对于高级白领，黑人初级白领与白人初级白领的收入差异。因为 $BLOCC_2$ 的系数为正，所以黑人在初级白领与高级白领中的收入差异比白人少 1501.20 美元，或者说在黑人中，该收入差异是-773.70美元而白人是-2274.90美元。同样，黑人高级白领与白人高级白领之间的期望收入差异为 B_1（-3793.3），而黑人与白人在初级白领工作的期望收入差异为 $B_1 + B_9$（$-3793.3 + 1501.2$），B_9 估计了黑人相对于白人在初级白

领与高级白领间的收入差异影响，－3793.3 为黑人与白人在高级白领上收入的差异，而－2292.1 为黑人与白人在初级白领收入上的差异。因此，黑人各职业的期望收入差异需要由两个回归系数捕捉，即 $\beta_j + \beta_{jk}$，其中 β_j 为职业虚拟变量（OCC_2 到 OCC_6），β_{jk} 为乘积变量的系数。我们可以把 β_j 和 β_{jk} 之间的关系定义为如下：

$$\beta_j = E(Y_i \mid_{\text{WHITE, OCC}j}) - E(Y_i \mid_{\text{WHITE, OCCref}}) \quad [4.1]$$

$$\beta_{jk} = [E(Y_i \mid_{\text{BLACK, OCC}j}) - E(Y_i \mid_{\text{BLACK, OCCref}})] - [E(Y_i \mid_{\text{WHITE, OCC}j}) - E(Y_i \mid_{\text{WHITE, OCCref}})] \quad [4.2]$$

$$= [E(Y_i \mid_{\text{BLACK, OCC}j}) - E(Y_i \mid_{\text{BLACK, OCCref}})] - \beta_j \quad [4.3]$$

因此，

$$\beta_j + \beta_{jk} = E(Y_i \mid_{\text{BLACK, OCC}j}) - E(Y_i \mid_{\text{BLACK, OCCref}}) \quad [4.4]$$

就像我们从方程 4.2 中看到的一样，乘积项回归系数的 t 检验不是黑人在不同职业中的期望收入净差异。我们所检验的假设要验证的是，对于黑人和白人，在指定职业类别与参照组间的收入净差异是否一样。

如果这些乘积项系数为负，我们就有证据说明，相对于白人，高级白领与其他职业类别的收入差异在黑人里更大。如果再加上 OCC_2 到 OCC_6 的负系数，那么黑人在各职业类别中的收入差异与白人相比就更加显著。但是，从回归结果看出，乘积项的系数是正值，那么在不同职业类别间，白人的收入差异相对而言就更显著了，因此，相比之下，黑人就没那么显著，或者说其各职业类别间的收入就更接近。事实上，黑人在不同职业间的收入可能没什么区别，而且当职业等级逐渐降低时，收入差异不论在黑人之间还是白人之间似乎都

是缩小的。

现在，我们进一步检验这些结论。为了证明之前的两个结论，我们必须确定当控制了教育和工作任期之后，非洲裔美国人在不同职业类别中的收入差异是否显著。可是没有一个计算机程序可以提供这种 t 检验，问题仍无法解决。交互项的 t 检验值只会告诉我们，从事不同职业的净效应对黑人和白人是否有显著不同。然而，知道作为一个技术工人、一个操作工人、一个服务业工作者或劳工相对于高级白领的种族差异，并不能确定对于黑人，从事某一职业类别的净效应，及其作为一个可靠的预测因子一定比另一个职业类别好。要回答这个问题，我们则需再深入一点。

表 4.4 不同种族所得收入的职业净效应

	白 人	黑 人
初级白领	−2274.9 (280.2)	−773.7 (776.7)
技术工人	−2418.4 (232.7)	−92.2 (683.5)
操作工人	−3427.2 (256.3)	−442.4 (646.5)
服务业工作者	−4513.4 (372.5)	−985.4 (682.6)
劳工	−4202.8 (399.0)	−818.9 (667.6)

表 4.4 通过控制教育和工作任期，检验了黑人和白人的职业取向效应。这里，同白人有关的项与表 4.2 中 OCC_2、OCC_3、OCC_4、OCC_5、OCC_6 的系数一样，这些系数估计的均是白人各职业类别与参照组——白人高级白领间

的收入差异。检验这些系数的统计显著性是为了了解在所有人群中，白人工作者在各职业类别中的收入差异是否存在。另外，如果我们要检验白人操作工人平均收入是否比白人技术工人低，那么，我们就应控制种族、教育和工作任期，通过方程3.1，模型5中的$B_3=-2418.1$、$B_4=-3427.2$及它们各自的方差、协方差，来比较这两个职业类别的收入差异。

对于黑人，一个职业类别的效应由两个回归系数捕捉：职业虚拟变量的回归系数和交互项的系数。例如，当要确定黑人初级白领和高级白领在期望收入上如何不同时，我们会将白人高级白领和白人初级白领的期望收入差异B_2和黑人相对于白人在高级白领和初级白领中的收入差异B_9加起来比较。根据表4.3，我们可以看出，要计算黑人初级白领的期望收入，就要通过计算黑人高级白领期望收入的两个回归系数(B_0+B_1)和另外两个回归系数(B_2+B_9)。[7]因此，对于黑人，表4.4中一个职业类别中的估测收入效应可以通过对适当的回归系数求和来构造。

正如我们在表4.4中看到的一样，黑人相对于白人，在非高级白领职业中显示的收入劣势要小一些。与白人的职业间收入差异不同的是，存在于黑人中最小的差异是高级白领与技术工人之间的差异，而不是高级白领与初级白领之间的差异。事实上，黑人高级白领与初级白领的收入净差异和高级白领与劳工的差异差不多。然而，这些关于黑人职业间的期望收入净差异仅基于回归系数。那么，回归检验可以使我们对这些估测的差异更自信吗？如果不可以，那么这些差异是否大多归因于抽样误差？

表4.2中乘积项系数的t检验告诉我们，除了初级白领，任意指定职业类别和高级白领在黑人或白人中的收入差异都是显著不同的。但是，我们还没有直接检验黑人职业间收入差异是否显著。那么，必要的t检验是估计两个回归系数的和相对于该统计样本分布的标准差的比率，即$(B_2+B_9)/SE(B_2+B_9)$。下面这个与方程3.1相似的方程正好可以达到此目的：

$$t=B_j+B_{jk}/[\mathrm{var}(B_j)+\mathrm{var}(B_{jk})+2\mathrm{cov}(B_j, B_{jk})]^{1/2} \quad [4.5]$$

与方程4.1到方程4.4一样，B_j代表第j个职业类的虚拟变量的回归系数，B_{jk}表示第k个自变量，在该例中，即为黑人与第j个职业虚拟变量乘积项的回归系数。t检验可通过将表4.4中与黑人相关的系数的代入方程4.5计算得出。我们发现，所计算出的t值没有一个是超过临界值±1.96的，从而可得出结论，剔除了教育年限和工作任期的影响后，尽管职业定位对白人期望收入水平有显著的影响，却没有真正提高黑人的平均期望收入。[8]

至此，我们再回到图3.1。我们将通过假设不同种族中的职业类别效应是不同的来开始本章的内容。就这张图，我们可以设想截距所体现的种族差异在各职业类别中是不同的。通过检验，我们发现，结果确实如此。对比图4.1与图3.1，可以更明显地看出，不同职业类别的收入差异强度对黑人和白人确实不同。

与之前相比，我们在表4.5列出了对图3.1和图4.1中12个组群的估计截距值。左边两列为通过模型4的估计值

计算出来的截距(见图 3.1),右边两列为通过模型 5 的估计值计算出来的截距(见图 4.1)。因为职业类别间种族的差异效应不同,所以通过模型 5 所得到的截距不仅修改了组群间的距离,还沿着 Y 轴重新排列了组群。我们需要谨记在心的是,这些组群的期望收入均是建立在受访者的教育年限和工作任期均为 0 的情况下的。

表 4.5　种族和职业的截距差异

	模型 4		模型 5	
	白　人	黑　人	白　人	黑　人
高级白领	5761.1	4573.0	5794.8	2001.5
初级白领	3445.0	2256.9	3519.9	1227.8
技术工人	3417.4	2229.3	3376.4	1909.3
操作工人	2594.5	1406.4	2367.6	1559.1
服务业工作者	1842.6	654.5	1281.4	1016.1
劳工	2154.3	966.2	1592.0	1182.6

从图 4.1 可以看出,白人和黑人在高级白领与技术工人之间的垂直距离都是一样的。对于这两个职业类别,代表技术工人的线比代表高级白领的线低 2344 美元。2344 美元为表 3.3 中列出的 OCC_3 的回归系数,其估计了在年收入上,作为一个技术工人相对于高级白领所带来的影响的大小。然而,在图 4.1 中,高级白领与技术工人的比较描述了不同职业类别中的种族效应。对于白人,这段垂直距离为 2418 美元;对于黑人,这段距离仅为 92 美元。这些数据来自表 4.2 中模型 5 的回归结果,并可直接从表 4.4 中获得。

尽管图 3.1 与图 4.1 截距之间的距离不同,但是所有组群的教育年限和工作任期的偏斜率都是一样的,即实线都是互相平行的,这表明图中所用的是教育的平均效应;所有虚

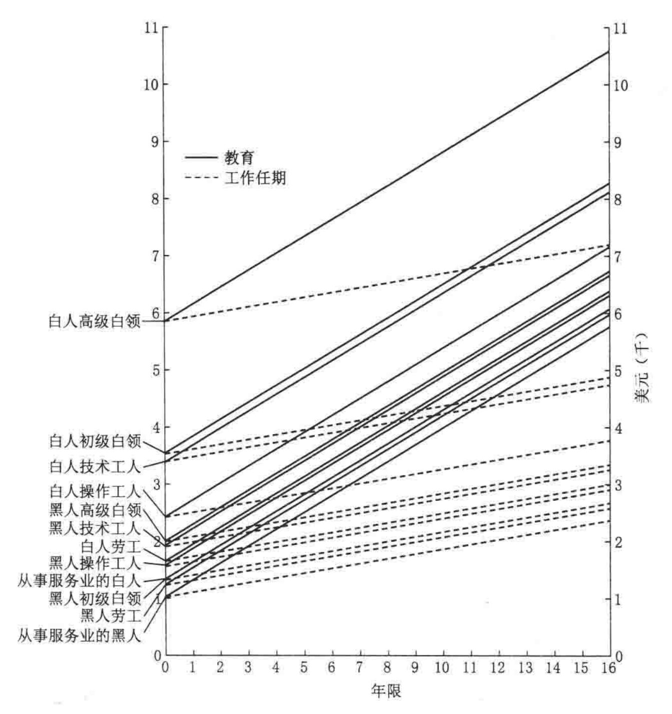


图 4.1 模型 5 的回归结果

线也是互相平行的，说明了工作任期也是由其平均效应表现的。本部分构建的最后一个模型对这点提出了质疑：对于所有组群，教育年限和工作任期的偏效应真的相同吗？对于这点，我们可以假设教育和工作任期的种族效应，然后再检验教育年限和工作任期对黑人和白人是否都一样。尽管我们发现每增加一年教育年限或工作任期，所获得的收入增加值对黑人和白人是不同的，但是一些线仍然呈现平行状态。因为有六组线是与黑人相关的，其中每一条线对应一个职业类别，还有六组线是与白人相关的，其中每一条线同样对应一个职业类别。我们发现，教育对于种族的差异效应表明，与

黑人工作者相关的实线是互相平行的，与白人工作者相关的实线也是互相平行的，但是这两个斜率可能不尽相同。[9]工作任期的也存在类似的情况。

为了检验假设 $\beta_{EDUC(whites)} = \beta_{EDUC(blacks)}$ 和 $\beta_{TENURE(whites)} = \beta_{TENURE(blacks)}$，我们构建了一个新模型来检验黑人和白人之间的差异效应。该模型基于模型 5，只是在模型设定中又新加入了两个变量。由于要检验的是变量关系的差异性，因此我们仍会用到交互项。在该模型下，新增加的交互项分别为 EDUC 和 BLACK 的乘积及 TENURE 和 BLACK 的乘积，分别记做 BLEDUC 和 BLTEN。对于样本中的黑人，BLEDUC 的分布应该和 EDUC 的分布是一样的。然而，对于样本中的白人，其在 BLEDUC 中全被编码为 0；对于 BLTEN 与 TENURE，情况与教育类似。该新模型可描述为：

$$
\begin{aligned}
\text{模型 6}: Y_i &= f(\text{种族，职业，教育，工作任期}) \\
&= \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 + \beta_4 \text{OCC}_4 \\
&\quad + \beta_5 \text{OCC}_5 + \beta_6 \text{OCC}_6 + \beta_7 \text{EDUC} + \beta_8 \text{TENURE} \\
&\quad + \beta_9 \text{BLOCC}_2 + \beta_{10} \text{BLOCC}_3 + \beta_{11} \text{BLOCC}_4 \\
&\quad + \beta_{12} \text{BLOCC}_5 + \beta_{13} \text{BLOCC}_6 + \beta_{14} \text{BLEDUC} \\
&\quad + \beta_{15} \text{BLTENURE} + u_i
\end{aligned}
$$

模型回归结果在表 4.6 中列出。除了 EDUC 和 TENURE，其他与模型 5 中相同变量的回归系数的解释方法基本不变。有一点需明确的是，估计这些效应基于其不仅控制了其他自变量，还控制了种族在教育年限和工作任期上的差异效应。

表 4.6　模型 6 的回归结果

		黑人的效应[a]
常数	4962.5 (435.9)	
BLACK	−1667.3 (901.3)	
OCC_2	−2155.4 (281.7)	−1068.8 (779.5)
OCC_3	−2167.9 (242.5)	−718.4 (703.4)
OCC_4	−3132.1 (268.8)	−1144.5 (672.7)
OCC_5	−4281.2 (378.9)	−1605.1 (703.4)
OCC_6	−3851.3 (411.0)	−1611.5 (703.7)
EDUC	359.1 (29.4)	186.3 (37.3)
TENURE	80.3 (7.6)	94.6 (13.1)
$BLOCC_2$	1086.5 (829.0)	
$BLOCC_3$	1449.5 (744.4)	
$BLOCC_4$	1987.6 (724.7)	
$BLOCC_5$	2676.1 (799.5)	
$BLOCC_6$	2239.7 (815.2)	
BLEDUC	−172.7 (47.5)	
BLTEN	14.2 (15.2)	
R^2	0.32434	
F	102.25	
R^2 的增量($R_3^2-R_1^2$)	0.00296	
变化过的 F	7.01***	

注：括号里为回归系数的标准误。

a. 黑人的系数是由加和模型 6 回归系数得出的；其标准误通过 $[\mathrm{var}(B_i)+\mathrm{var}(B_j)+2\mathrm{cov}(B_iB_j)]^{1/2}$ 计算得出。

*** 表示系数在 0.001 显著水平下显著。

尽管对于大多数变量，其系数估计和显著性检验都相同，但是相对于模型 5，模型 6 有一些回归结果的变化需提及。尤其是虚拟变量黑人的系数减少了不止一半，而且其在 0.05 显著水平上也不显著了。另外，交互项$BLOCC_3$ 也明显变小了，而且在 0.05 水平下边际显著。我们应该如何解释这些变化呢?

为了回答这个问题，我们必须看看新加入的这两个变量——教育和工作任期的交互项。BLTENURE 的回归系数统计上并不显著，这表明在一个雇主的情况下，每增加一年的工作任期，其黑人或白人增加的收入是一样的。在这个发现下，我们回想从模型 4 到模型 5，当第一次加入了乘积项后，从对 OCC_2 到 OCC_6 的系数的解释方法可以运用到该模型对 TENURE 的回归系数的解释中。在这里，TENURE 的回归系数表明，为同一雇主工作每增加一年，白人工作者的收入每年会增加 80.30 美元。BLTENURE 回归系数估计了黑人与白人的工作任期净差异为 14.20 美元。那么，我们就可以知道，黑人工作者为一个雇主工作每增加一年，其收入每年会增加 94.50 美元。然而很明显，BLTENURE 的标准误与回归系数大小相近，这说明了该效应很弱。因此，我们可以得出如下结论：工作任期对黑人和白人的影响基本一样。

教育的情况就有所不同。EDUC 的系数告诉我们，当控制了其他变量的影响之后，每增加一年的教育年限，白人的收入可增加 359.10 美元。BLEDUC 回归系数表明，每增加一年的教育年限，黑人的收入仅会增加 186.40 美元(359.10 美元－172.70 美元)。BLEDUC 的显著性检验表明，其在统

计上是显著的。对于所有人群，若不考虑其他变量，那么，每增加一年的教育年限，黑人所增加的工资相对于白人会少一些。因为黑人和白人的教育净效应不同，所以模型 5 所用的黑人和白人的平均教育效应其实低估了每增加一年教育年限白人的教育回报，同时高估了每增加一年教育年限黑人的教育回报。

在描述有关种族—职业的交互项时，我们可以对这些影响效应问同样的问题。尤其当我们已经知道，工作任期对黑人的净效应与白人没有显著不同，但是教育的确不同，其对白人的净效应要大于黑人。我们不知道教育年限是否会显著影响黑人的期望收入水平。要回答这个问题，我们必须回到方程 4.4，通过检验有关的回归系数的加和值来估计黑人的教育净效应。将相关系数代入方程 4.4 后，我们发现：

$$
\begin{aligned}
t &= 359.1 + (-172.7)/[(862.691) \\
&\quad + (2253.569) + 2(-862.691)]^{1/2} \\
&= 186.4/37.29 = 5.00
\end{aligned}
$$

教育确实会影响白人和黑人的期望收入，然而，黑人的平均教育回报率比白人低。

对于之前提出的问题——我们该如何考虑 BLACK 和 $BLOCC_3$ 的影响在方向上的逆转呢？现在我们可以提供这样一个解释：同模型 5，当教育对收入的净效应没有种族差异时，则黑人的回归系数告诉我们，在高级白领中，黑人有着非常显著的收入劣势；在控制了其他变量后，黑人高级白领平均比白人高级白领少挣4000美元。同样，对于技术工人与高级白领之间的期望收入净差异，白人(2418.40美元)也比黑

人(92.20 美元)多很多。这些都可以从图 4.1 中看出。然而,一旦考虑了模型 6 中种族对教育的差异效应时,就会发现在高级白领中,黑人和白人的期望收入净差异在双尾检验的 0.05 显著性水平下不显著了。尽管从模型 5 到模型 6,虚拟变量黑人的回归系数的标准差增加了,但是最重要的变化应该是系数本身的点估计。模型 6 中 BLACK 的回归系数比模型 5 中的一半还小,这个现象反映了黑人高级白领与白人高级白领的 Y 截距的差异变小了。这点我们可以在图 4.2 中看出。但是,当我们沿着横坐标轴向高教育年限方向移动

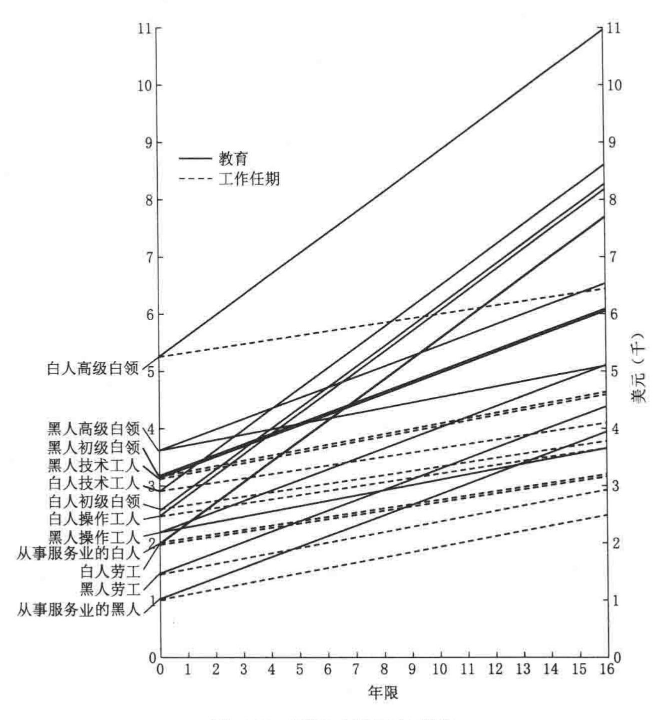


图 4.2 模型 6 的回归结果

时,会发现黑人高级白领与白人高级白领间的距离变大,这说明白人高级白领相对于黑人高级白领的相对收入优势随着教育年限的提高而变大。从表 4.6 的回归系数我们可以算出,当教育年限为 0 时,白人高级白领与黑人高级白领的收入差异为 1667 美元;当教育年限变为 16 年时,白人高级白领的期望收入比黑人高级白领高 4430 美元。

当我们把注意力转移到 $BLOCC_3$ 时,会发现 $BLOCC_3$ 在模型 6 中的回归系数比模型 5 中的小,因此,我们不可以拒绝有关"当控制了模型中其他因素和教育的种族效应后,技术人员的净效应在黑人和白人之间相同"的零假设。但通过模型 5 和模型 6 的对比,可以知道,白人高级白领与黑人高级白领之间的期望收入净差异部分可归因于获得比较高教育等级的黑人所挣得的收入,总是不如同等条件下的白人多。如果额外的教育年限带来的收入差异被允许,换句话说,一旦我们可以承认额外的教育年限所引起的收入增加在白人中比黑人多,我们就可以解释为什么黑人高级白领获得的期望收入比白人高级白领少,因为他们额外教育的回报率相对白人而言很低。我们不想把最初的观测结果——黑人高级白领相比白人高级白领有收入劣势——视为无效;相反,根据黑人的收入劣势,模型 6 的结果提出了一个可能的解释。另外,我们知道,在技术工人中,那些最初看上去由种族差异导致的收入差异效应,其实部分也归因于在该职业类别中,黑人相对于白人所累积的教育回报差异。

第2节 对各组群分别进行回归

我们已知年纪较大的黑人工作者与白人工作者的期望收入有差异，现在，需要通过控制职业、教育年限、工作任期来研究该现象，进而一步步扩展模型设定，把有关白人与黑人中的额外变量的差异效应检验也包括在内。该步骤可能会引起读者的疑问，在允许自变量在每个组群中有所不同的情况下，为什么我们会用模型6对整个样本进行估计，而不是对每个组群分别进行回归估计呢？为什么不把样本分成黑人和白人两组，用期望收入对职业虚拟变量、教育年限和工作任期对每个组群分别进行回归？事实上，若是检验假设和标准OLS假设都可以通过恰当的统计过程得到满足，那么，这些方法都是等价的。

为了构建一个含有交互项的全样本模型，我们要注意以下六点：

第一，当没有乘积项或交互项时，自变量的系数告诉我们的是一个“平均效应”，而当其他自变量也包含在规范里时，则为“平均偏效应”。

第二，当把乘积项加入到模型设定中时，我们可以通过比较两个模型的 R^2 值来确定是否要用各组群的平均效应来提高模型拟合度。如果 R^2 的增量是由于加入了乘积项后而变

得足够大,我们就可以拒绝零假设(各组群的效应是相同的)。

第三,当我们用模型 6 对全样本进行估计时,从 OCC_2 到 OCC_6 的回归系数的 t 检验可以测量白人工作者的职业净效应,而 BLACK 的回归系数的 t 检验则可以解释非洲裔美国人在高级白领中的期望收入净效应。

第四,为了检验一个自变量的效应是否对黑人工作者显著,我们必须构建两回归系数的加和的 t 检验。

第五,为了检验两个自变量的效应是不是显著不同,例如,检验操作工人是不是和服务业工作者不同,我们必须构建两回归系数差异的 t 检验。

第六,要对乘积项进行 t 检验,我们需要确定解释变量的效应是否因种族差异而不同。

通过对每个组群分别进行回归,例如,先对黑人进行回归,再对白人进行回归,我们就可以自动估计出不同组群的效应。换句话说,对每个组群,如果要知道一个指定自变量是否有显著效应,就必须考虑上述第四点是否必要。然而,如果估计不同组群的回归的目的是评定组群差异效应的显著程度,那么,该差异效应检验就非常必要了。

当缺乏详尽的检验时,研究者可能会陷入两难。想象一下,当要分析政治激进主义为年龄的函数时,我们期望得到的是年龄和政治激进主义的关系随教育程度的不同而不同,尤其是当我们假设年龄和政治激进主义的关系是否与大学毕业有关时。从大学毕业组与非大学毕业组分别随机选取 500 个样本,然后分别对其进行回归,我们可能会发现一个明显的年龄效应“差异”。例如,假设年龄对大学毕业生的影响为－0.16,而对非大学毕业生的影响为－0.32。那么,因为

大学毕业生的回归系数是非大学毕业生的一半，我们就可以说年龄对大学毕业生的影响比那些没有接受大学教育的人小吗？大多数读者在此可能会意识到这么说有风险。如果该研究的目的是从样本推及整体，那么同时考虑点估计和误差是非常必要的。当要评估回归模型中自变量影响的重要性时，就不是简单的系数强度的问题了，而是系数强度相对于标准差的问题。当要描述有关相对影响的强度，即这些影响是不是相等时，也需要考虑该问题。估计回归系数的差异强度必须相对于差异的标准误。在这种情况下，由于信息非常有限，我们很难判断在受教育程度不同的组群里，年龄对政治激进主义的相对影响。

表 4.7 对组群分别回归的结果

	黑 人	白 人
常数	3295.2	4962.5
	(416.0)	(494.9)
OCC_2	−1068.8	−2155.4
	(411.2)	(319.8)
OCC_3	−718.4	−2167.9
	(371.1)	(275.3)
OCC_4	−1144.5	−3132.1
	(354.9)	(305.2)
OCC_5	−1605.1	−4281.2
	(371.2)	(430.2)
OCC_6	−1611.5	−3851.3
	(371.2)	(466.6)
EDUC	186.3	359.1
	(19.7)	(33.3)
TENURE	94.5	80.3
	(6.9)	(8.6)
R^2	0.31887	0.24450
F	61.06	105.50
RSS/$(n-k-1)$	4046652.4	18754709.2
N	921	2290

注：括号里为回归系数的标准误。

现在我们假设，除了回归系数估计，我们已经知道了这些估计的标准误：回归系数−0.16的标准误为0.11，系数0.32的标准误为0.14。那么我们会发现，大学毕业生组的影响在常用的显著性水平下并不显著。[10]此时，是不是就可以下结论说，年龄的效应在非大学毕业生组更大呢？答案当然是否定的。尽管我们已经证明在非大学毕业生组中，年龄是估计政治激进程度的一个重要预测因子，但是我们并没有对该差异效应进行进一步的推理。如果研究问题是：两种影响是不是相等？或者说 $H_0: \beta_{CG} - \beta_{NCG} = 0$ 是否成立（其中，β_{CG}是年龄在大学毕业生组的影响，β_{NCG}为年龄在非大学毕业生组的影响）？那么，此时就不再是 β_{CG}或 β_{NCG}本身相对于我们的零假设是不是显著，比较合适的估计应该是用 $\beta_{CG} - \beta_{NCG}$。该统计强度必须是相对于 $\beta_{CG} - \beta_{NCG}$ 的标准差来估计的。当回归系数都来自同一个方程时，我们可以用方程4.5来进行检验。但是，如果系数是从不同的回归中估计出来的，该检验的定义就会有些不同。[11]

要说清楚这些问题，我们必须先回到收入数据。然后对白人工作者和黑人工作者分别进行回归，其回归结果可在表4.7中看到。由于回归系数是在不同的样本中估计出来的，那么，所有 B（如 B_1、B_2、B_3等）差异影响估计必然是不相关的。那就是说，协方差的估计为0。因此，方程中的标准差变为根号下方差的和。然而，这个不同组群的方差系数部分是基于整体方差的组群估计。总体方差因为总体方差的每两个估计只是建立在“部分”样本上的，所以其用的是残差平方和的不同部分。这就表明了通过合并两个组群的信息计算总体方差的合并估计值的必要性（Kmenta，1986）。另外，由

于组群的大小可能不同(例如,白人的数量可能比黑人多两倍以上),这时,该合并估计值必须通过合适的自由度来对每个组群估计加权(Long & Miethe, 1988)。

假设各组群的方差是相等的(也就是我们所说的方差的同质性),那么,计算总体方差的合并估计值的方程为:

$$s_{pooled}^2 = \frac{(n_1 - k_1 - 1)s_1^2 + (n_2 - k_2 - 1)s_2^2}{N - (k_1 + k_2 + 2)} \qquad [4.6]$$

其中,n_1 和 n_2 是组群中的事件数,$N = n_1 + n_2$;k_1 和 k_2 为每个组群中包含的自变量的数目;s_1^2 和 s_2^2 是组群各自回归出的平均残差平方和。[12] 用来检验组群回归系数差异的 t 检验公式为:

$$t = \frac{B_1 - B_2}{s_{pooled}\left(\frac{s_{B_1}^2}{s_1^2} + \frac{s_{B_2}^2}{s_2^2}\right)^{1/2}} \qquad [4.7]$$

其中,$s_{B_1}^2$ 和 $s_{B_2}^2$ 分别为 B_1 和 B_2 的方差,s_1^2 和 s_2^2 与上式相同。通过执行上述 t 检验,我们即可重现在全样本回归模型中的交互项的 t 检验。例如,将表 4.7 中的回归结果代入方程 4.6 及方程 4.7,从而检验对黑人和白人教育的影响是不是都一样。那么,为了计算总体方差的合并估计,我们有:

$$s_{pooled}^2 = \frac{(921 - 8)(4046652) + (2290 - 8)(18754709)}{3211 - 16}$$

$$= 14551750$$

将其带入 t 检验方程后,我们发现:

$$t = \frac{186.3 - 359.1}{3814.7\left(\frac{386.8}{4046652} + \frac{1108.9}{18754709}\right)^{1/2}} = \frac{-172.7}{47.5} = -3.6$$

通过 t 统计量分子、分母的比较,连同表 4.6 里所有回归系数和 BLEDUC 的标准差,我们可以对这两个过程的等价性看得更加清楚。然而,在证明该等价性时,我们仅简单地把相关系数带入方程 4.6 和方程 4.7,而没有检查是否满足了隐藏在 t 检验背后的假设。事实上,直到现在,所有的讨论都忽略了是否符合 OLS 假设的问题。在构建一个含有二进制编码的虚拟变量回归系数的解释过程中,我们一直坚持 OLS 假设是没有争议的,不管事实如何。现在是时候纠正这种错误了。确实,由于当前我们把重心放在推理检验上,这就需要我们更仔细地检验这些假设,特别是在已经有了先例后,在对有关从全样本或者组群回归中出现的差异效应下任何结论前,我们必须检验方差的同质性(或称"同方差性")。

然而,尽管是否符合 OLS 假设非常重要,但从表 4.7 和表 4.6 的比较结果来看,这样做还是会有问题。表 4.6 中列出的黑人职业净效应的显著性检验是基于模型 6 的全样本估计,当其分别对黑人与白人回归后,无法得到重现。虽然对回归系数的估计没有分别,但是标准差却明显不同。当对各组群分别进行回归估计时,相比表 4.6 中的结果,黑人的标准差变小了,而白人的标准差变大了。那么,我们如何解释这种不一致性呢?

第 3 节 | 处理异方差性

我们已经根据隐含的假设(收入水平和收入结构的决定因素与种族有关并随种族不同而变化)发展了模型。换句话说,我们开始只是识别黑人和白人的收入水平总差异,之后发展到可以解决解释变量差异效应的问题,比如职业虚拟变量和与教育年限、工作任期有关的交互项等。然而,这些检验都存在潜在问题,因为其假设是基于两组群的同方差性。

尽管从模型 6 中用全样本数据获得的回归系数估计和分别回归组群数据得出的估计一样,但是由全样本回归得来的总体方差 RSS/($N-k-1$) 只有在同方差性的假设被满足后,才基本上与分别回归组群得来的方差合并估计值相等。由组群回归带来的平均残差平方和对于黑人是 4046652,而对白人为 18754709。相对于对每个组群分别回归估计——分别只估计了其中一部分观测(该例中,分别为 921 个黑人和 2290 个白人),模型 6 的全样本估计(14551750)用了所有的观测值,因此全样本估计是建立在最多的信息上的。但是,只要我们假设 u_i 对所有的 X_{ki} 都是个常数,那么从全样本数据得来的估计即对总体方差的无偏估计。但是对于本例,“u_i 的方差不会因组的不同而不同”这一假设有效吗? 要回答这个问题,必须更仔细地研究我们的回归结果。

同方差性的假设指出，给定特定值的自变量干扰项的条件方差为一个常数，即 σ^2。对于比较不同组群的研究问题，检验其异方差性是很有必要的。许多异方差检验都要求研究者检验 OLS 估计的残差平方，即 e_i^2。基本的统计教材也会为该问题及其检验过程提供比较概述性的讨论（Gujarati, 1988; Johnston, 1984）。根据我们的侧重点，我们将会把注意力主要集中到两个检验和一个讨论上。

在此例中，我们假设方差的大小是种族的函数，即方差在不同的种族中是不同的。一个直接的检验是可以比较从黑人和白人的分别回归中得到的平均残差平方和。如果同方差性存在，那么两组群在估计回归平面的误差方差上也应该相同。因此，我们可以提出零假设，即这些方差都是相等的，其又可表达为，$H_0: \sigma_1^2 = \sigma_2^2$，同时可用平均残差和作为这些参数的估计。因此我们构建了一个有关相对方差的比率，那么当比率为 1 时，方差相等；当比率偏离 1 时，同方差假设就站不住脚了。假设 u_i 服从正态分布且同方差性的假设成立，那么该比率遵循 F 分布。从而，我们可以构造如下的检验：

$$F_{n_1-k_1-1,\ n_2-k_2-1} = \frac{\mathrm{RSS}_1/n_1-k_1-1}{\mathrm{RSS}_2/n_2-k_2-1} \qquad [4.8]$$

其中，分子由较大方差的组群回归得出，而分母由较小方差的组群回归得出。在该情形下，k_1 和 k_2（回归中所包含的自变量个数）是相等的，因其模型设定相同。将表 4.7 的结果代入，我们有：

$$F_{9\ 132\ 282} = 18754709.2/4046652.4 = 4.63$$

对于该样本量大小的组群，F 值在 0.001 显著水平下为 4.63，这表明我们要拒绝同方差的零假设，而倾向于异方差性。对于模型 6，其回归系数的估计仍然是无偏的。然而，在异方差下，t 检验是不准确的。而且，在组群间的异方差条件下阐述解释变量的组群差异效应非常复杂，因为影响差异的回归检验并不清楚。我们并不知道导致该检验结果的是组群间的差异效应，还是组群间的方差差异。

当方差不等时，检验正态分布的均值等价性问题即著名的Behrens-Fisher 问题（Amemiya，1986：36）。针对该问题的解决方法有很多。这些方法主要依靠一些数据转换或者再加权来处理异方差问题，或者通过重新计算检验统计量的分布来调整偏差。在本例中，异方差性的问题似乎比较容易解决。

最初，我们从传统回归模型设定的基础上着手，这样，解释变量的备择组合就有被假定的可能性，但是我们不会质疑用来检验这些关系的模型设定。虽然已经用 INCOME（以工资和奖金挣来的美元数衡量）作为因变量，但是实际所用的模型设定可能并不是最佳选择。例如，如果工资呈对数的正态分布（该断言已经为大量的经济学文献所支持），那么异方差性可能就会在模型设定的误差中产生。

第 4 节 | 解释半对数方程的虚拟变量

我们通常使用自变量和因变量原来的度量来解释回归方程。根据此惯例,我们保留了读者已非常熟悉的回归系数解释方法,即期望值 Y 随 X 的单位变化而变化。有时,函数规范会要求自变量或因变量,或者两者一起变换。研究文献中最常见的变换就是对数变换。当变量呈高度偏斜分布时,这种变换就尤其有用。[13] 尽管对数变换通常是处理异方差性的补救方法(Gujarati, 1988; Maddala, 1992),但在本例中,将其转变成收入的自然对数是为了澄清一个对收入分布的理解问题,即 5000 美元与 10000 美元的收入差异和 50000 美元与 55000 美元的收入差异的意义是不同的。而在初始的收入分布度量中,不管该增量在分布中是如何计算的,5000 美元的差异就是 5000 美元的差异,它始终只是一个增量而已。相反,如果把增量 5000 美元视为比例项,我们就可以看出,在初始5000美元的基础上再增加 5000 美元,相当于增加了 100%,而在初始 50000 美元的基础上增加 5000 美元,相当于只增加了 10%。那么,如果要说出与 50000 美元成比例的同等效力,则应该是在其基础上增加到100000美元,或者说增加 50000 美元,这样才增加了 100%。通过收入的对数变化,我们可以用比例项来表示自变量和因变量之间

的关系。在一个半对数模型中，只能通过将因变量或者自变量转换为对数形式来实现；在本例中，我们所用的是因变量收入的对数变换模式，而保持自变量的度量标准不变。

我们可以将 Y 的半对数模型定义如下：

$$
\begin{aligned}
\text{模型 7: } \ln(Y_i) &= f(\text{种族,职业,教育,工作任期}) \\
&= \beta_0 + \beta_1 \mathrm{BLACK} + \beta_2 \mathrm{OCC}_2 + \beta_3 \mathrm{OCC}_3 \\
&\quad + \beta_4 \mathrm{OCC}_4 + \beta_5 \mathrm{OCC}_5 + \beta_6 \mathrm{OCC}_6 + \beta_7 \mathrm{EDUC} \\
&\quad + \beta_8 \mathrm{TENURE} + \beta_9 \mathrm{BLOCC}_2 \\
&\quad + \beta_{10} \mathrm{BLOCC}_3 + \beta_{11} \mathrm{BLOCC}_4 \\
&\quad + \beta_{12} \mathrm{BLOCC}_5 + \beta_{13} \mathrm{BLOCC}_6 \\
&\quad + \beta_{14} \mathrm{BLEDUC} + \beta_{15} \mathrm{BLTENURE} + u_i
\end{aligned}
$$

当 X_{ki} 为一个连续变量时，我们将其回归系数解释为在给定的 X 的绝对变化下，Y 的相对变化。例如，TENURE 变化一年，相应的 Y 的成比例变化多少。如果我们用 100 乘以 β_8，那么，我们得到的是 X 一个单位的绝对变化所带来的收入的百分比变化。例如，如果 B_8 为 0.014，我们可以说每增加一年的工作任期，所预测的 Y 会增加 1.4%。该解释在自变量为连续度量时是有效的，但是 Halvorsen 和 Palmquist(1980)告诉我们，对于虚拟变量，其回归系数不可以这样解释。

假设 X_{ki} 是一个虚拟变量，由于虚拟变量用的是离散编码，如 0 和 1，则不可以通过回归系数来定义斜率。因此，对于虚拟变量，我们不能将其回归系数作为因变量对虚拟变量的导数。另外，虚拟变量的回归系数通过因变量的单位变化捕捉的是指定组群与参照组之间的均值差异。当 $Y^* = \ln Y$ 时，半对数模型设定中的虚拟变量的回归系数表达了 $\ln Y$ 的

相对变化。就如 Halvorsen 和 Palmquist(1980)所说的，半对数回归中的虚拟变量系数等于：

$$B_1 = \ln \frac{1+\hat{Y}_1-\hat{Y}_{ref}}{\hat{Y}_{ref}} \qquad [4.9]$$

其中，$\hat{Y}_1$ 为被编码成 1 的组群的 $\hat{Y}$ 期望值，$\hat{Y}_{ref}$ 为参照组的 Y 的期望值。为了用 Y 的原始度量标准而不是相对于其对数变换模式的分布表达虚拟变量对 Y 的百分比效应，我们就需要用对数函数的反函数，即指数形式或者反对数形式。此时，编码为 1 的组群(而不是参照组)的百分比差异可表示为：

$$100[\exp(B_1)-1] \qquad [4.10]$$

因此，若虚拟变量的系数，如黑人的回归系数 B_1 为－0.632，不难算出，用 1 减去以 e 为底，－0.632 为指数的值为－0.468，这表明了黑人的期望值 Y 比参照组白人的期望值低 46.8%。

表 4.8 的左边列出了模型 7 的全样本估计结果；右边是对黑人和白人分别进行回归的结果。

我们第一个任务是观察现在是否满足同方差的假设。将相关系数代入方程 4.8 后，我们有：

$$F_{9\,132\,282} = 0.27774/0.21983 = 1.263$$

根据所得的 F 值，我们无法拒绝同方差的零假设，因此，现在可以着手检验黑人和白人之间的差异效应了。

在总结重要的结论之前，我们先来比较一下表 4.8 里两个回归模式的结果。首先，我们可以发现，表的两部分的系数估计是相同的。其次，如果再仔细观察会发现，不同回归

表 4.8 所有解释变量和交互项对 LN(INCOME)的效应

	全样本回归		组群样本回归	
	模 型	黑 人[a]	白 人	黑 人
常数	8.353 (0.056)		8.353 (0.054)	7.720 (0.109)
BLACK	−0.632 (0.115)			
OCC_2	−0.244 (0.036)	−0.029 (0.099)	−0.244 (0.035)	−0.029 (0.108)
OCC_3	−0.174 (0.031)	0.056 (0.090)	−0.174 (0.030)	0.056 (0.097)
OCC_4	−0.328 (0.035)	−0.056 (0.086)	−0.328 (0.033)	−0.056 (0.093)
OCC_5	−0.585 (0.049)	−0.105 (0.090)	−0.585 (0.047)	−0.105 (0.097)
OCC_6	−0.510 (0.053)	−0.209 (0.090)	−0.510 (0.051)	−0.209 (0.097)
EDUC	0.049 (0.004)	0.043 (0.005)	0.049 (0.004)	0.043 (0.005)
TENURE	0.014 (0.001)	0.027 (0.002)	0.014 (0.001)	0.027 (0.002)
$BLOCC_2$	0.215 (0.106)			
$BLOCC_3$	0.230 (0.095)			
$BLOCC_4$	0.272 (0.093)			
$BLOCC_5$	0.480 (0.102)			
$BLOCC_6$	0.301 (0.104)			
BLEDUC	−0.006 (0.006)			
BLTEN	0.013 (0.002)			
RSS 的平均值	0.23654		0.21983	0.27774
R^2	0.42489		0.32987	0.30082

注:括号里为回归系数的标准误。

a. 黑人的系数是由加和模型 6 的回归系数得出的;其标准误通过 $[\mathrm{var}(B_i)+\mathrm{var}(B_j)+2\mathrm{cov}(B_iB_j)]^{1/2}$ 计算得出。

模式所得到的标准差也非常接近。第三,通过这两个模式的回归检验,我们可以得到相同的结论:对于白人,在控制教育和工作任期的情况下,高级白领有着非常显著的收入优势,同时,教育年限和工作任期对期望收入有正的净效应;对于高级白领,黑人没有白人那样的收入优势,事实上,当控制了模型中的其他因素时,高级白领的期望收入比劳工多很多,但是其他的职业差异却没有那么显著。教育年限所带来的收入净效应对黑人和白人来说没什么差别,然而,工作任期所带来的效应对黑人比对白人大。此外,种族和职业的交互项告诉我们,职业的净效应因种族的不同而不同,职业差异导致的期望收入差异在黑人中表现甚小。

我们可以通过描述黑人在各职业中的净效应来总结这些结果:黑人高级白领相对于白人高级白领有明显收入劣势,这种劣势基本上在除了服务业工作者以外的所有职业类别中都存在。在服务业工作者中,控制了教育和工作任期后,黑人和白人之间的期望收入差异并不显著。[14]

第5节 | 检验两组以上的异方差性

关于检验异方差性的方法及文献有很多。例如，Goldfeld-Quandt 检验，其在观测量不是很大之时是一个比较合适而且常用的方法。然而，此方法需要把观测分成两组（Goldfeld & Quandt, 1972; Gujarati, 1988）。另一个常被引用的检验过程是 Glejser（1969）提出的，他认为要检验一系列回归，就要证明$|e_i|$是模型中每个自变量的函数（其局限性可参见 Gujarati, 1988; Maddala, 1992）。在该例中，我们先用模型6的回归结果来计算 e_i，然后用$|e_i|$对 BLACK 进行回归。此方程的 F 检验是为了说明是否要拒绝误差项的同方差性。从两组扩展到 j 组要求$|e_i|$对 $j-1$ 个虚拟变量进行回归，此时，F 检验可决定是否拒绝每组方差都是相同的假设。

作为对已估计的回归模型的回归诊断。对自己是否遇到异方差感兴趣的研究者，往往需要在做回归分析前就进行检验。然而最近有关同方差性检验的比较分析的结果显示，这些检验的统计强度和稳健性差距非常大（Conover, Johnson & Johnson, 1981）。这些检验的一个共同局限就是它们对非正态的分布非常敏感。就此，Levene（1960）提出了一个在比较分析中较好的检验，该检验的结构其实与 Glejser 的异方差性检验非常类似。Levene 提出用单向方差分析分析

绝对偏差值时，如果用偏差的中位数代替偏差均值，其稳健性检验会明显提高。为了进行该检验，研究者必须在一开始就计算好$|Y_{ij}-\tilde{Y}_j|$，其中，$\tilde{Y}_j$ 表示第 j 组中位数。因为单向方差分析等价于对 $j-1$ 个虚拟变量进行虚拟变量回归，所以我们就可以估计回归方程：

$$|Y_{ij}-\tilde{Y}_j| = B_0 + B_1 D_1 + \cdots + B_{j-1} D_{j-1} + e_i$$

该方程的 F 检验决定了同方差性零假设是否应该被拒绝。

第 6 节 | 用非独立检验进行多重比较的方法

通过讨论有关估计组群差异的方法，我们找到了另一个值得讨论的话题——多重比较的问题。该问题同时是统计学推论中的一个论点(Miller, 1966)，源于对单系列的估计进行多组比较。要进行比较的数目越大，则至少有一个比较是显著的可能性也越大。现在关于如何进行多重比较的显著性检验仍然有许多争议。这里，我会简单介绍两种方法——Bonferroni 检验和 Fisher 的 t 保护方法(Darlington, 1990:249—275)。

在本例中，我们检验了不同职业类别的期望收入差异。将高级白领作为参照组，表明我们会直接比较五对职业类别，即其他五类职业相对于参照组的差异。然而，我们也引入了 t 检验，它可以检验回归系数之间的差异。我们知道，所有可能的对比数目共有 $j(j-1)$ 个，其中 j 表示类别的数目。然而，若比较的先后顺序无关紧要，可能的对比数目会减少一半。由此，对于六个职业组，我们可以生成 15 个可能的成对比较。用这些方法来阐述该问题，前提是必须明确这些可能的比较是否相互独立。对于该例，高级白领与初级白领的比较和高级白领与技术工人的比较不是相互独立的，因为对收入较高的高级白领的概率选择可能会影响这两对比较结

果。从而,高级白领与初级白领、技术工人与操作工人的对比才是独立的比较,或者说是正交的比较。但是,对这些比较的显著性检验不是完全独立的。原因在于在计算这两对比较的标准差时,我们用的都是总体方差的估计值,即 RSS/$(N-k-1)$,所以该估计的随机波动会影响两个检验的 t 值。

Ryan(1960)发展了对 Bonferroni 不等式和独立检验的应用,他证明了Boferroni不等式为大多数修正过的非独立检验的显著性水平提供了略为保守的估计。该不等式的目的是为了根据多重比较的实际情况,提供一个修正后的显著性水平。Bonferroni 检验类似于修正显著性水平(CSL),其通过把检测得出的所有结果的最显著 t 值所对应的概率 p 与被检验的结果个数(即 Bonferroni 校正因子或者 BCF)相乘,从而可以计算出一个修正显著性水平。因此,若 15 个结果中最显著的 t 值对应的 p 值为 0.003,那么对该估计修正显著性水平即 $\text{CSL}=15(0.003)=0.045$。其中,15 就代表 BCF。如果要用第二小的 p 值来估计系数的显著性,我们就应该将该 p 值乘以 14。该方法被称为“压条法”,这个过程一直到找到第一个非显著结果才会停止。

Dunn(1961)告诉我们,对于双尾检验,表达式 $[1-(1-\text{CSL})^{\text{BCF}}]$ 可为我们提供一个修正显著性水平的上限。然而,Bonferroni 检验有时太过保守,Dunn 检验有时又显得太过自由。如果想更好地理解 Bonferroni 不等式的用法,我们必须考虑两个极端情况(Darlington, 1990)。第一个情况为两个检验在 -1 上相关。例如,如果我们对 B_k 进行双尾检验,则要检验两个零假设:$\beta_k \geqslant 0$ 和 $\beta_k \leqslant 0$。这两个检验在 -1 上相关,因为要拒绝第一个零假设,则必须排除第二个零假设。

用 Bonferroni 不等式，我们可以说，至少一个零假设在0.025显著性水平被拒绝的可能性不会比 0.05 大，即 2×(0.025)。在该情形下，p 值和修正显著性水平 0.05(即一个双尾检验零假设 $\beta_k = 0$)的显著性水平一样大。因此，这么说来，Boferroni 不等式并不保守。第二个情况即两个检验在＋1上相关。假设我们有一个包含 j 个类别的变量，我们每次都用第一个类别去和其他任何一个类别比较，这样就可以产生 $j-1$ 个比较。该情况在我们有一个 j 类的定性变量并且在回归模型里含有根据该变量产生的 $j-1$ 个虚拟变量时便会发生。如果所有除去参照组的类别都有无限的样本量和同样的均值，那么所有比较的 t 值也会一样，此时，拒绝一个事件的零假设就意味着其他的零假设也要被拒绝。在这种情况下，修正显著性水平就等于我们观测到的 p 值。因此，如果我们要用 Bonferroni 不定式，我们就要高估修正显著性水平。因此，Bonferroni 公式的不准确性与各检验之间的相关性有关。各检验间的正相关越大，误差就越大。

相比之下，Fisher 的方法就更自由一些。运用该方法时，研究者进行 F 检验来检验各类别之间没有不同的零假设是否要被拒绝。如果 F 检验结果显著，那么研究者就可以进行任意类别的比较，因为包含在这些对比中的 t 检验已被从 F 检验得出的显著性结果证实了。在讨论各种模型的回归结果中，第一步总是检验当加入一系列虚拟变量来代表一个定性特征(如职业的一系列虚拟变量)或者一系列交互项(如种族和职业的乘积项)后，所出现的 R^2 的增量的统计显著性。在通过 F 检验建立统计显著性后，我们就可以用 Fisher 方法来进行多重比较了。



第5章

可替代虚拟变量编码方案

迄今我们所接触到的虚拟变量大都采用二进制编码并指定单一参照组，其实还有很多其他的编码方案也是可行的。比如，效果编码和对比编码这两种替代性的方案。同样，这两种方法要求我们用 $j-1$ 个虚拟变量来表示具有 j 个类别的名义变量。

第1节 | 效果编码虚拟变量

如第2章提到的，有些研究者倾向于选择一个中间类别作为参照组，而不是直接用一组按序数分布的极端类别，这样的选择可以解释为，通过建立组群比较来模拟指定类别和所有样本的"平均值"的差异。若想对组群和样本平均进行比较，效果编码的解释结构将比二进制编码更方便。

表5.1 虚拟变量的效果编码和对比编码

职业类别	效果编码				
	E_1	E_2	E_3	E_4	E_5
高级白领	−1	−1	−1	−1	−1
初级白领	1	0	0	0	0
技术工人	0	1	0	0	0
操作工人	0	0	1	0	0
服务业工作者	0	0	0	1	0
劳工	0	0	0	0	1
职业类别	对比编码				
	C_1	C_2	C_3	C_4	C_5
高级白领	0.5	1	0	0	0
初级白领	0.5	−1	0	0	0
技术工人	−0.25	0	0.5	1	0
操作工人	−0.25	0	0.5	−1	0
服务业工作者	−0.25	0	−0.5	0	1
劳工	−0.25	0	−0.5	0	−1

为了方便用二进制编码和其他编码方案进行比较，我们将继续将收入作为因变量，将种族和职业作为名义自变量。表 5.1 举例说明了我们如何分别利用效果编码和二进制编码虚拟变量来捕捉种族和职业类别的信息。表格的上半部分描述了通过效果编码产生的五个虚拟变量，其保留了高级白领作为参照组。然而，我们知道在二进制虚拟变量编码中，参照组通常被编码为 0，而对于通过效果编码产生虚拟变量的参照组，习惯编码为－1。每一个虚拟变量所捕捉的职业类别比较即参照组和编码为 1 的组。在本例中，E_1 为高级白领和初级白领的对比；E_2 为高级白领和技术工人的对比；E_3 为高级白领和操作工人的对比等等。如果需要比较的组群样本大小一样，编码为 0 的组就不会影响比较结果。但是，如果组群的样本大小不一致（这是常见的例子），编码为 0 的组群所带来的影响便会出现，尽管其很小。实际上，编码为 0 的组群的影响随着所有编码为 0 的观测值的偏离样本均值的程度增大而增大（Cohen & Cohen, 1983）。

表 5.2 记录了效果编码虚拟变量和收入的零阶相关性、均值及标准偏差。除了职业类别的虚拟变量，还有一个种族虚拟变量 ERACE，其为白人时编码为 1，为黑人时编码为 0。二进制编码虚拟变量均值等同于指定组群中的事件比例，而效果编码虚拟变量的均值则指出了参照组（编码为－1）和其他编码为 1 的组群的大小差异。实际上，均值即 $(n_j - n_{ref})/N$。例如，E_1 将 644 个高级白领编码为－1，337 个初级白领编码为 1，其他为 0。那么，E_1 的均值为 $(337-644)/3211=-0.096$。负号表示参照组比编码为 1 的组群具有更多的观测值；强度表示这个差异的大小相对于总体样本的大小。这

表 5.2 虚拟变量的均值、标准差和相关性

	效果编码的虚拟变量					
	E_1	E_2	E_3	E_4	E_5	收入
ERACE	−0.132 (0.000)	−0.057 (0.001)	−0.231 (0.000)	−0.272 (0.000)	−0.333 (0.000)	0.313 (0.000)
E_1		0.563 (0.000)	0.565 (0.000)	0.662 (0.000)	0.644 (0.000)	−0.270 (0.000)
E_2			0.444 (0.000)	0.584 (0.000)	0.560 (0.000)	−0.242 (0.000)
E_3				0.586 (0.000)	0.562 (0.000)	−0.354 (0.000)
E_4					0.660 (0.000)	−0.403 (0.000)
E_5						−0.420 (0.000)
平均值 标准差	−0.096 (0.544)	0.052 (0.671)	0.045 (0.666)	−0.111 (0.527)	−0.093 (0.547)	6903.220 (4629.954)
	对比编码的虚拟变量					
	C_1	C_2	C_3	C_4	C_5	收入
CRACE	0.271 (0.000)	0.132 (0.000)	0.172 (0.000)	0.164 (0.000)	0.088 (0.000)	0.313 (0.000)
C_1		0.265 (0.000)	−0.257 (0.000)	−0.006 (0.358)	0.027 (0.063)	0.396 (0.000)
C_2			−0.068 (0.000)	−0.002 (0.462)	0.007 (0.343)	0.270 (0.000)
C_3				0.009 (0.310)	0.068 (0.000)	0.056 (0.001)
C_4					0.000 (0.491)	0.105 (0.000)
C_5						0.040 (0.012)
平均值 标准差	−0.021 (0.346)	0.096 (0.544)	0.150 (0.389)	0.007 (0.706)	−0.018 (0.443)	6903.220 (4629.954)

注:括号里为相关系数的概率值。

组均值指出，高级白领的观测值数量超过了初级白领、服务业工作者和劳工，但少于技术工人和操作工人。效果编码虚拟变量的方差是两个待比较组群之间相对频数的函数，即 $p_j + p_{\text{ref}} - (p_j - p_{\text{ref}})^2$。那么，如上所述，$E_1$ 的方差是高级白领和初级白领之间相对频数的函数，即 $s_{E_1}^2 = 0.1050 + 0.2006 - (-0.0956)^2 = 0.2965$，$s_{E_1} = 0.544$。

由 E_1 到 E_5 与收入的相关性系数可以看出，高级白领和体力劳动者之间的对比（即 E_5）是最强烈的，因为其均值偏离最大。但是，由于样本在各个职业和种族组群之间不是均匀分布的，所以零阶相关性的解释还不太明确。在这里，虚拟变量间的零阶相关性仍然表示所对应组群的相对大小。对于效果编码的虚拟变量，相关系数 0.50 只有在各个组群具有相同的样本量时才会出现。当参照组样本大于其组群时，零阶相关性系数会大于 0.50（例如，E_1 和 E_5 的或 E_1 和 E_4 的相关性）；当参照组样本量少于其他组群时，零阶相关性系数会降至 0.50 以下。

回归结果

虽然不同的编码方案会使回归系数在数值上有所不同，但是总体模型拟合度（由 R^2 表示）以及种族和职业类别虚拟变量对收入的影响的显著性（由模型 1 的 R^2 的 F 检验和模型 3 相对模型 1 的 R^2 增量指出）仍可重现第 3 章中模型 1 和模型 3 的结果。不同的编码方案会影响信息的捕捉方式，即对于不同编码方案，其组群间差异方式的排列不同，但是不会影响总体的结果，因为组群之间的结果差异相

对于前面的估计保持不变，我们只是从另外一个角度来观察而已。

之前提到过，二进制编码虚拟变量只有在其他虚拟变量被控制的时候才能对参照组和指定组进行比较。换句话说，组群比较只能作为部分效果而存在。效果编码虚拟变量的情况是相似的，尽管比较的本质有所改变。这里，在控制了 E_2 到 E_5 后，E_1 的偏回归系数表达了初级白领与样本中所有组群的比较结果。这样，由偏回归系数估计得出的数值就等于第 j 组与没有加权的所有组群的收入期望值差异，即 $B_k = \overline{Y}_j - \sum \overline{Y}_j / j$，其中，$j$ 是原来名义度量的类别数，$\overline{Y}_j$ 是第 j 类组群的均值。我们可通过截距得到没有加权过的所有组群均值的均值，该值可以作为所有组群差异的计算参考点。

没有加权的均值和总体样本的均值的度量不同，其数值相同与否，取决于组群均值相对于组群大小的变异性。总体的样本均值可以看成加权后的所有组群均值的均值，因为我们计算样本均值时，是把每组的均值与该组的事件数相乘后求和，再除以总样本量所得出的。计算没有加权的均值的平均值，相当于给每一组样本赋予相同的权重 1，不管该组中的事件数为多少。这样做的结果是，一些只包含少量事件且度量不太准确的组群均值会与有大量观测值且度量较准确的组群均值得到同等对待，然而这种准确度上的差异会在回归系数的标准差中反映出来。该过程也可以使非标准化的回归系数独立于组群的相对大小。

表 5.3 用不同编码方案的回归结果

	效果编码			对比编码	
	模型 1	模型 3		模型 2	模型 3
常数	6220.5 (85.8)	6277.8 (83.2)		6567.7 (78.8)	6751.5 (88.7)
种族	1601.4 (85.8)	838.0 (86.2)			838.0 (86.2)
E_1		853.5 (196.9)	C_1	5247.6 (223.4)	6443.8 (260.3)
E_2		129.2 (141.4)	C_2	1510.6 (137.2)	2842.1 (271.1)
E_3		−908.9 (140.9)	C_3	1987.3 (192.4)	1494.8 (196.3)
E_4		−1817.1 (211.4)	C_4	695.5 (102.1)	519.1 (102.3)
E_5		−1952.2 (200.6)	C_5	172.2 (163.1)	67.5 (161.1)
R^2	0.09792	0.24624		0.22400	0.24624
F	348.3	174.4		185.0	174.4

注:括号里为回归系数的标准误。

现在,我们来考虑表 5.3 中模型 1 的回归结果。效果编码的种族变量(ERACE)是模型中唯一的自变量,所以,其截距等于黑人的收入均值加上白人的收入均值,然后除以 2。读者可以从表 2.2 的数字中证明这个数值。$B_{ERACE} = 1601.4$ 是白人收入期望值(7821 美元)和没有加权过的白人和黑人的平均值的差异,即斜率。

模型 3 包含了 ERACE 和职业虚拟变量。从 ERACE 的回归系数可以看出,一旦考虑了不同职业类别的收入差异,白人的收入正效应仍然显而易见。同样,当控制了黑人/白人的收入差异后,初级白领和技术工人平均比其他组群的期望收入高,且初级白领的优势超过技术工人。操作工人、服

务业工作者和劳工均在均值以下，其中服务业工作者和劳工的收入劣势最大。在控制其他职业虚拟变量后，效果编码虚拟变量、收入与一个职业虚拟变量的偏回归系数可以被理解为该组的“反常”或者“独特”的性质度量（Cohen & Cohen，1983）。通过对半偏相关系数取平方，我们可以评估每一个类别的特性使收入变异的程度。

如前几章所述，我们可以用每一特定组群的编码数乘以该组虚拟变量的估计系数来预测该组群的收入。如果虚拟变量是二进制编码的，该过程就可叙述为把虚拟变量的回归系数加入其所代表的组群，并丢弃所有被编码为0的虚拟变量。用效果编码虚拟变量，参照组统一被编码为－1，例如，高级白领在所有职业虚拟变量中就被编码为－1。因此，要计算白人高级白领的预测收入，我们有：

$$
\begin{aligned}
Y_{UWC} &= 6277.8 + 838(1) + 853.5(-1) + 129.2(-1) \\
&\quad - 908.9(-1) - 1817.1(-1) - 1952.2(-1) \\
&= 10811.3
\end{aligned}
$$

同样，如果要计算黑人高级白领的期望收入，我们只需将ERACE(838)的系数乘以－1，其他不变，即可得出黑人预期收入为9135.3。通过对比表5.3中模型3和表3.2中模型3的期望收入值，读者可以证实，不论我们对虚拟变量使用哪种编码方案，所得出的期望收入值都是一样的。因此，效果编码虚拟变量和二进制编码虚拟变量的主要区别在于其参考点的定义。不同于估计每个组群相对某个特定参照组的差异，效果编码是把每个组群与所有组群进行比较。

Suits(1983)证明了，不论这些差异的解释如何，通过在

二进制编码虚拟变量的估计回归系数中加一个常数，就有可能变更解释框架，从而可以通过各组群间没有加权过的均值解释所有组群的偏差。我们可以考虑最简单的情况，即用收入对一个二进制编码的虚拟变量种族(BLACK)进行回归后产生的结果，如表 3.1 模型 1 中所列出的：

$$Y_i = 7821.9 - 3202.9(\text{BLACK}) + e_i$$

通过一个常数 c 对 B_k 进行调整，我们可以把解释框架从二进制编码转化为效果编码模式。我们可以通过 $\sum(B_k + c) = 0$ 来确定 c 的值，其中 B_k 为二进制编码虚拟变量的回归系数，那么，$c = -(\sum B_k)/j$，其中 j 为定性度量的类别个数。在该例中，c 等于$-(-3202.9/2)=1601.45$。通过对每个虚拟变量的回归系数加一个 c，在常数项里减去一个 c，我们有：

$$Y_i = 6220.45 - 1601.45(\text{BLACK}) + 1601.45(\text{WHITE})$$

这样就可以表达 WHITE 的回归系数，尽管它在原先的模型设定中为参照组。我们可以简单地假设 WHITE 的效应在原先的规范中为 0。

当有多于一个的定性变量被加入规范中时，调整可以在虚拟变量组群内确定。比如，当用 INCOME 对 BLACK 和职业虚拟变量进行回归后，我们有：

$$\begin{aligned} Y_i = 10811.4 &- 1676(\text{BLACK}) - 2842.1(\text{OCC}_2) \\ &- 3566.4(\text{OCC}_3) - 4604.5(\text{OCC}_4) - 5512.7(\text{OCC}_5) \\ &- 5647.8(\text{OCC}_6) + e_i \end{aligned}$$

c_{RACE}像之前那样确定，即 $c_{\text{RACE}} = -(-1676/2) = 838$。$c_{\text{OCC}} = -[(-2842.1)+(-3566.4)+(-4604.5)+(-5512.7)+$

(−5647.8)]/6=3695.58。我们把 c_{RACE}加到种族的两个类别里，其中 WHITE 的回归系数根据原先的规范被设置成了 0，然后我们再把 c_{OCC}加到职业的六个类别里，其中高级白领的回归系数在原先的规范中也被设置成 0，之后我们从常数项中减去c_{RACE}和 c_{OCC}。这样，我们就可以表达任意组群相对于没有加权的所有组群的平均值的影响了。

$$
\begin{aligned}
Y_i = & 6277.8 - 838(\text{BLACK}) + 838(\text{WHITE}) \\
& + 3695.6(\text{OCC}_1) + 853.5(\text{OCC}_2) \\
& + 129.18(\text{OCC}_3) - 908.9(\text{OCC}_4) \\
& - 1817.12(\text{OCC}_5) - 1952.2(\text{OCC}_6) + e_i
\end{aligned}
$$

第2节 | 对比编码虚拟变量

表5.1下半部分列出了一系列对比编码的虚拟变量。在三个条件下，研究者可以通过对比编码指定其感兴趣的比较：(1)对于含有 j 类的名义变量，其表达需要指定 $j-1$ 个对比；(2)任何对比编码虚拟变量的编码的加和必须为0；(3)任何两个虚拟变量的编码必须正交。根据经验产生的对比编码需要我们最初就将一系列类别归入两个合并的组群中。

在此例中，我们可以区分白领和蓝领工作者。C_1 定义了所有白领与所有蓝领的比较。因为白领组合并了高级白领和初级白领，因此每个类别被编码为0.5。同样，由于蓝领包含了其他四个组，所以每个类别被编码为-0.25，负号表示蓝领与白领的对比。其中，0.25是四个组群加入相等权重后产生的聚合群的结果，该四个编码之和为1。

剩余的虚拟变量在它们最初的分类中定义了对比。比如，C_2 对比了两个白领组的成员，因为每个组都是独立的，所以一组编码为1，另一组为-1。C_3 比较的是技术工人和操作工人与服务业工作者和劳工之间的区别，前两组编码为0.5，后两组编码为-0.5。C_4 定义了技术工人和操作工人之间的区别，C_5 比较了服务业工作者和劳工之间的区别。[15]我们可以将连续两对编码的乘积求和来检验该对比编码组的独立

性。例如，对 C_1 和 C_2 的编码乘积求和，我们有 (0.5)(1) + (0.5)(−1) + (−0.25)(0) + (−0.25)(0) + (−0.25)(0) + (−0.25)(0) = 0。

表 5.2 的下半部分列出了表 5.1 中定义的对比编码虚拟变量的零阶相关性、均值和标准差。其中，该对比编码虚拟变量的均值和标准差是组群相对大小的函数，但是由于编码规范中包括的数值均小于 1，因此这些频数和均值之间的关系对解释的用处不大。

如前一个例子，这一系列虚拟变量呈现出和其他变量的零阶相关性，即便之前它是用来定义组群对比而现在有了正交的性质。但是要求编码正交的条件与要求变量正交不同，对比编码虚拟变量之间的相关性是各组群相对大小的函数。只有当观测值在各组中均匀分布时，相关性才为 0。[16]

有关对比编码虚拟变量和收入之间的零阶相关性的解释，确实不甚明确。如 C_2、C_4 和 C_5，其解释基本上和效果编码虚拟变量一样，因为对于这三个变量，对比是由一个编码为−1 的组、一个编码为+1 的组和其他编码为 0 的组而捕捉到的。这种编码方法和效果编码差不多，除了一点不同，即在对比编码虚拟变量中，−1 不是被分配到所有相同的组中的。因此，这些相关性是用来衡量平均收入编码为−1 和 1 的组间的差异程度，当然是在考虑了 INCOME 的方差的情况下。然而，如果组群大小不相等，编码为 0 的组也可以在其度量上影射该信息。

解释只有两种数值的编码的虚拟变量相关性(如 C_1)非常直接，因为 C_1 把白领编码为 0.5，把蓝领编码成−0.25，收入和 C_1 之间的零阶相关系数的平方测量了收入方差的比

例，该比例可由白领和蓝领之间的差异来解释。当其他虚拟变量没有被控制的时候，编码方案的目的是要计算加权后的各组群均值的均值。例如，由于高级白领和初级白领的编码是一样的，所以和零阶相关性有关的白领平均 INCOME 是一个合并的均值，忽略了原先细分的组群信息。换句话说，白领的均值是包括了所有白领工作者，即高级白领和初级白领加在一起的收入均值。因此，零阶相关性包括了加权后相似编码组的均值。我们知道，该数值可以通过每组的均值与每组的事件数相乘求和后，再除以所有事件数得出。然而，必须指出的是，该解释只有在处理仅有两个可能数值的对比编码虚拟变量的零阶相关性时才合适。所以，我们可以得出一个普遍的结论：尽管对比编码虚拟变量基于所有变量，为我们提供了总结回归结果的一个有用的替代方法，但是这种一次只能比较一个变量的简单描述性统计并不特别有用。

回归结果

表 5.3 的右边两栏列出了模型 2 和模型 3 的回归估计结果。没有包括模型 1 的估计结果，是因为它和效果编码虚拟变量完全一样。再提一下，第 3 章表 3.1 中列出的 R^2 和 F 检验可重现，这强调了三种方法在多元回归分析中的等价性。

包含在回归估计中的剔除过程相对二元测量，仍需要一个更直接的系数解释方法，尽管一些额外计算是必需的。对模型截距的解释和对模型效果编码虚拟变量的解释相同，即没有加权的所有组群均值的均值，它为我们提供了估计组群

效应的参考点。每个虚拟变量指出了两个组群或组群集合的对比。变量的偏回归系数是没有加权的组群均值的均值与用来创建对比的编码差异的函数。由于剔除过程,编码为0的组已被排除,不进行比较。

组群对比 C_j 被定义为:

$$C_j = B\frac{n_{g1}+n_{g2}}{(n_{g1})(n_{g2})} \qquad [5.1]$$

其中,n_{g1} 为第一个集合中所包含的组群数,n_{g2} 为第二个集合中所包含的组群数,B 是虚拟变量的回归系数。例如,系数 C_1 是白领和蓝领对比的函数,将其带入方程5.1,我们得到:

$$C_1 = 5247.6\frac{2+4}{(2)(4)} = 5247.6(0.75) = 3935.7$$

其他对比的计算如下:

$$C_2 = 1510.6(2) = 3021.2$$

$$C_3 = 1987.3(1) = 1987.3$$

$$C_4 = 695.5(2) = 1391.0$$

$$C_5 = 172.2(2) = 344.4$$

通过表2.2列出的组群均值,读者可以证明,这些对比确实可以重现组群均值差异或者没有加权的组均值的均值。

这些对比的标准差可通过将回归系数的标准差乘以一个我们之前用于加权系数的因子而得到。例如,高级白领和初级白领的对比的标准差(C_2)为 $(137.2)(2) = 274.4$。该值和表3.1中模型2列出的 OCC_2 的数值一样。同时,回归系数的 t 检验使我们可以估计由虚拟变量定义的对比是否可以推广到整体。在模型2中,回归系数 C_1 到 C_4 在0.001的水

平上都很显著，但 C_5 却不是。从而我们可以得出结论：总体而言，没有加权过的高级白领和初级白领的平均收入比蓝领的平均收入高；高级白领平均收入比初级白领高；没有加权的技术工人和操作工人组比服务业工作者和劳工组的平均收入高，但是劳工的期望收入却与服务业工作者没有显著差别。[17]

当所有其他虚拟变量被控制时，半偏回归系数的平方表示了可由一个特定对比所解释的 Y 的样本方差的比例。我们来看看表 5.3 中模型 2 的半偏系数，假设其最大部分的方差来自白领和蓝领之间的对比（例如，$0.365^2=13.3\%$）。然而，半偏相关系数的平方和不能为方程提供 R^2 值，因为 C 变量是相关的。只有当所有的组群大小都一样时，对比编码虚拟变量之间才是无关的，而且只有当回归因子正交时，半偏相关系数的平方和才等于 R^2。

第6章

虚拟变量用法专题

至此,我们一直在用同一个数据集的同系列变量来探索逐渐复杂化的模型中虚拟变量的解释方法。在这一过程中我们也发现,引入虚拟变量可以使建模更加灵活。除了那些之前提到的假设,虚拟变量还经常用于一些其他形式。本章我们将探究一些虚拟变量在回归分析中的其他使用方法。

第1节｜logit模型中的虚拟变量

越来越多的研究者在logistic回归模型中运用二分类或多分类的因变量。由于许多研究问题涉及组群差异，在logistic模型中，虚拟变量回归因子已很常见。假设一个模型以死亡率为因变量、性别为虚拟变量，那么logit回归的虚拟变量的回归系数代表一个和性别相关的死亡对数优比(log-odds)的增量或者减量。然而，解释一个对数优比并不像解释一个简单的机会比率(odds ratio)那么有吸引力。那么，我们可以转换解释框架吗？答案是肯定的。我们可以通过logit系数的反对数来完成从相加效应(我们会在用线性相加模型预测几率对数时详细说明)到乘积效果，即把因变量变为一个简单的机会比率(Alba，1988)。由于对数转变可以使我们用相加的形式代表相乘关系，因此这种从相加关系到相乘关系的转变会伴随着从对数优比到机会比率的转变。

为了解释这个方法，我们先来看看Idler和Kasl(1991)的研究结果，即上了年纪的女性的预期死亡率作为其自身健康和健康风险因素评估的函数。对于该模型，因变量为四年里死亡的对数优比。如果受访者去世了，则编码为1；若没有去世，则编码为0。主观的健康状态由一组三个虚拟变量来衡量，以“健康状况非常好”为参照组。健康状况控制了包括

癌症、糖尿病、间歇性跛行和高血压，如果存在其中任一情况，则编码为 1。此外，还有需要协助的活动个数、日常不可自理的活动个数（ADL）、身体质量指数、年龄以及测量现在或者以前是否吸烟的两个虚拟变量，其中，以不吸烟者作为参照组。表 6.1 记录了有关 logistic 回归结果的估计。右边栏为虚拟变量 logit 系数的反对数。我们可以看出，当控制了其他变量时，健康状况差的相对于健康状况非常好的，其死亡的对数优比会增加。换句话说，在其他条件相等的情况下，健康状况差的女性的死亡几率是健康状况非常好的女性的 3.12 倍。同样，当控制了其他因素时，健康状况一般的女性的死亡几率是健康状况非常好的 2.85 倍，健康状态好

表 6.1 虚拟变量的 logistic 回归

	logistic 回归系数(*B*)	反对数(*B*)
常数	−6.308***	
自我评估健康状况		
差(同非常好比较)	1.138*	3.12
一般(同非常好比较)	1.047*	2.85
好(同非常好比较)	0.862*	2.37
糖尿病	0.963***	2.62
ADL 值	0.041	
活动	0.393*	
间歇性跛行	0.982	2.67
高血压	0.369	1.45
年龄	0.061**	
现在仍吸烟(同不吸烟比较)	0.769***	2.16
以前吸烟(同不吸烟比较)	−0.312	0.73
体重(kg)/身高^2(m)	−0.076***	

注：* 表示系数在 0.05 显著性水平上显著；
** 表示系数在 0.01 显著性水平上显著；
*** 表示系数在 0.001 显著性水平上显著。
资料来源：Idler 和 Kasl(1991)。

的是健康状况非常好的2.37倍。此外，吸烟的净效应告诉我们，现在还在抽烟的女性的死亡几率是不抽烟女性的 2.16 倍。对于样本成员，过去吸烟然后戒掉的女性的死亡几率比不抽烟的女性低，即为不吸烟女性的 0.73 倍，然而估计差异并不显著。

模型中的因变量是进行了对数转变的，其虚拟变量系数解释要分两部分说明。第一部分(见第 4 章后半部分)通过系数的算术转变，读者可以解释其中的百分比差异。在这一章中，我们对乘积效应作出了解释。表面上来看，虚拟变量的系数在第 4 章运用的半对数模型中的解释和 logit 模型解释不同。相反，这两种解释只在方差上有微小的差别。在方程 4.10 中，我们把相对影响定义为存在某种特征的百分比差异，其由编码为 1 的虚拟变量说明。例如，在白人中，当控制了模型中的其他变量后，劳工的期望收入比高级白领的收入少 40%。然而，如果不从回归系数的反对数里减 1，我们的结论是针对职业差异的乘积项的，即劳工的收入是高级白领的 60%。不管怎样，用虚拟变量回归系数估计的反对数转变模型来解释其相对影响是非常重要的。

第 2 节 | 非线性检验

虚拟变量常用来代表名义编码的自变量类别。但是，我们也会用虚拟变量来表示定序变量或者区间变量。例如，当我们怀疑自变量与因变量之间存在非单调性或者曲线性关系，但又没有很好的基础来预测这种特殊形式的曲线性时，虚拟变量回归提供了一个非常有用的替代方法来取代多项式回归或者算术转变模型。通过用一系列虚拟变量代表一个定量自变量，我们把整体分布分成几小段，然后检验虚拟变量之间是否存在线性或者曲线性的关系。

现在，我们用一个比较熟悉的例子来解释收入和教育之间的关系。与其断言教育的影响在整个区域都是相同的，不如先检验每增加一年的教育所带来的收入增量是否由其在整个分布中的位置而定。为了检验这个教育和收入之间的曲线性关系，我们要评估以下两个模型：

模型 6.1：$Y = f$(教育年限) $= \beta_0 + \beta_1 \mathrm{EDUC} + u_i$

模型 6.2：$Y = f$(代表各教育水平的虚拟变量)

$$= \beta_0^* + \sum \beta_j ED_j$$

在模型 6.1 中，教育(EDUC)是作为一个定量变量的，而在模型 6.2 中，教育被指定为一系列的虚拟变量，符号为 ED_j。其

中，当受访者没有接受过正式教育时，ED_0 编码为1；当受访者完成了超过一年的教育时，ED_1 编码为1；当受访者已经完成了17年的正式教育时，ED_{17}编码为1。参照组为接受了18年正式教育的受访者。

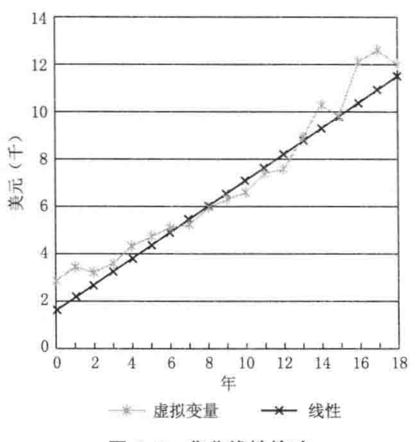


图6.1　非曲线性检验

图6.1对这两个模型结果作出了解释。乍看之下，会感觉其关系确实有些曲线化。在整个教育区域内，期望收入差异在连续的组群之间是不一致的。斜率在低教育水平上比较平缓，随着教育程度的升高，其呈现出比较陡的趋势，尤其是完成了12年教育之后。用虚拟变量代表教育年限的模型相对于线性模型，可解释更多的方差。模型6.1解释了19.6%的收入方差，而模型6.2解释了21.4%的收入方差。当检验放宽了线性回归假设后，要检验被解释方差的增量是否统计上显著，就需要如下的 F 检验，其中分子的自由度是由虚拟变量模型中多出的虚拟变量个数决定的，分母的自由度等于事件数减去虚拟变量模型中的参数数目：

$$F=\frac{(R_2^2-R_1^2)/(\mathrm{df}_2-\mathrm{df}_1)}{(1-R_2^2)/(N-\mathrm{df}_2)} \qquad [6.1]$$

将该 F 检验运用到模型 6.1 和模型 6.2 中，我们得到：

$$F_{17\ 319\ 2}=\frac{(0.21390-0.19624)/(19-2)}{(1-0.21390)/3192}$$

$$=\frac{0.00104}{0.00025}=4.16$$

当在 0.01 显著性水平下与 F 的临界值比较时，该 F 值是统计上显著的。

如果自变量在度量上是真正连续的，原先的变量必须在用一系列虚拟变量表示前，重新编码成离散的类别。当这一步骤成为必需时，第二个模型和现在文中出现的模型或多或少会有些不同，因为原先的定量自变量和虚拟变量是一起包含在模型设定中的，但是 F 检验的公式和两个方程的 R^2 相比是不变的。

第 3 节 | 分段线性回归

虚拟变量可以让我们对一个回归线上斜率的突变建模。当斜率是逐渐变化的时候，即当 Y_i 和 X_{1i} 在 X_{2i} 上以线性方式增加或减少时，我们就可以用一个乘积项来捕捉该效应。当斜率突变时，我们可以用虚拟变量来协助估计该变化的强度和显著性。当我们可以界定出定量自变量（X_i）分布的临界值，而且希望看到 X_i 和 Y_i 之间的关系在该临界值两边不同时，这个方法就相当有用。例如，零售业的营销人员经常可以从佣金中得到部分补偿，这些佣金与他们卖出的商品的数量有分级相关的关系。同样，投入和产出可能与经济规模有关，因此我们可根据该经济规模来修改某个产出水平下的投入方程。关于这个，我们会在后面具体说明。

假设我们有两个分布：第一个分布列出了全部产出；第二个分布列出了全部投入。我们可以进一步假设，当期望产量达到 5000 时，每单位投入会减少多少。因此，5000 即一个临界值 X^*。为了估计斜率，即每单位投入在 $X^* = 5000$ 时的变化，我们必须先通过临界值计算每个产出水平的偏差，即 $(X_i - X^*)$，然后定义一个虚拟变量（D_i），使其在产出超过临界值 5000 时为 1，否则为 0。那么，模型可以写成：

模型 6.3：$Y_i = B_0 + B_1 X_i + B_2 (X_i - X^*) D_i + e_i$

其中，B_1 估计的是产出不到 5000 时的斜率，$(B_1 + B_2)$ 估计的是产出超过5000时的斜率。因此，B_2 估计的是斜率的变化，B_2 的 t 检验提供给我们的是对斜率变化估计的显著性估测。

继续看我们的例子，当用投入对产出进行回归时，我们可以得到以下结果：

$$Y_i = \underset{(27.455)}{143.798} + \underset{(0.006)}{0.109}\text{OUTPUT} + e_i$$

该方程解释了 Y_i93.75%的方差，这说明，当产出增加一个单位值时，整体投入会增加大约 11 美分。换句话说，即每件物品的边际成本为 10.9 美分。当评估一个分段回归模型时，我们会发现：

$$Y_i = \underset{(34.264)}{87.059} + \underset{(0.010)}{0.129}\text{OUTPUT}$$
$$- \underset{(0.018)}{0.045}(\text{OUTPUT} - \text{OUTPUT}^*)(D_i) + e_i$$

该方程解释了整体投入 95%的方差。和之前的模型相比，它有 1.25%的显著提高。另外，我们知道，当产出少于 5000 时，每单位的边际成本大约是 13 美分。然而，当产量高于 5000 时，边际成本会降到 8.4 美分(0.129−0.045)。

第4节 | 时间序列数据中的虚拟变量

当数据为截面数据时,虚拟变量可以提供一个在因变量期望值下估计组群差异的方法。在这种情况下,组群可由我们期望与因变量的分布结构相关的特征来定义。当数据成时间序列分布时,虚拟变量可以对组群观测进行分类。然而,在时间序列数据里的分组更有可能被定义为相对重要的事件。就像截面数据一样,虚拟变量如同分布机制的代理,不仅很难衡量,而且也很复杂。由于它们起着代理的作用,因此对差异背后的机制解释很容易引起争议,那么,其有效性及对任何模型设定的解释都可以成为一个反对的理由。

虚拟变量在时间序列回归中可用来捕捉区域性的或者组群差异的信息。但是,它们也可以用于检验参数结构稳定性和构建季节指数的工具。例如,研究美国工会会员增长的人员通常认为,《瓦格纳法案》的通过是工会主义发展的主要原因;研究军方花费增长的调查人员发现,对战争动员的效应调整是非常必要的;还有一些建模研究试图说明,税法的变化是某项投资的盈利变化的决定性因素。我们知道,一个重要事件的影响可以导致趋势线的转变及其过程的结构调整。

关于结构稳定的例子,我们可以研究一下 Carl Chen

(1984)有关三里岛核事故对市场模型稳定性的影响问题。他的数据由1978年第一季度到1980年第一季度的70个公用事业股的周价格组成。这些核公司的股票在该事件后大幅度下降，Chen通过比较事件之前和之后的参数估计来检验市场模型的稳定性。根据需要，他提出了以下模型：

$$模型\ 6.4：r_{jt}=\beta_{j0}+\beta_{j1}r_{mt}+u_{jt}$$

其中，r_{jt}为j股在时间t内的每周回报，r_{mt}为以标准普尔指数表示的市场回报，u_{jt}是随机干扰项。

用虚拟变量来检验截距(β_{j0})和斜率(β_{j1})稳定性的方法需要我们把样本分成两个子样本期。在这种情况下，我们定义一个虚拟变量，当观测发生在事件之前时，$D=0$，发生在事件之后，则$D=1$，事件发生的那周从观测中删除。这样，检验模型变为：

$$模型\ 6.5：r_j=\beta_{j0}^*+\beta_{j1}^*r_m+\beta_{j2}^*D+\beta_{j3}^*r_mD+u_j$$

其中，β_{j2}估计了两个样本期截距值的差异，β_{j3}估计了两个样本期斜率系数的差异。研究核公司的组群，即那些在1980年有多余10%核燃料的结果如下，其中括号里的为t值：

$$\begin{aligned} r_j = &-0.0022-0.0031D+0.3553r_m+0.0614r_mD \\ &\quad(1.43)\qquad(1.32)\qquad\quad(4.13)\qquad\quad(0.50) \end{aligned}$$

根据这些结果，“没有结构性变化”的零假设不能被拒绝。

第5节 | 虚拟变量和自相关

我们可以考虑一个简单的时间序列模型,预测 Y_t 作为 X_t 和一个虚拟变量 $D(D=1$ 表示后两个时期)的函数,该模型用来估计包含两个时期跨度的过程。那些对估计 Y 水平的变化和形成 Y 的过程变化感兴趣的人,可能会试图估计以下模型:

$$Y_t = B_0 + B_1 D + B_2 X_t + B_3 DX_t + e_t$$

其中,B_1 估计了从第一期到第二期的变化,B_2 估计了第一期时 X 的效应,B_3 估计了在第二期和第一期之间,X 对 Y 的影响变化的差异。

在估计时间序列模型时,研究人员必须注意它是否违反了没有自相关的假设。如果残差检验告诉我们误差之间有相关性,则 OLS 估计是无效的。要处理自相关问题,研究者必须经常假设干扰项是从一阶自相关过程中产生的,也就是说,现时段干扰项是之前时段干扰项的函数。这种相关的程度由自相关系数 ρ 来测量。那么,补救措施就会涉及估算广义差分方程,即用 $(Y_t - \rho Y_{t-1})$ 对 $(X_t - \rho X_{t-1})$ 进行回归。其中,ρ 为对自相关系数的估计。但是虚拟变量呢?是不是要对它们进行同样的转变呢?其实不用(参见 Maddala,1992:

321—322)。假定虚拟变量定义了两组观测，最关键的观测为第二期中的第一个观测，研究人员必须对这些组群里的观测进行如下定义：

第一，对于所有第一期的观测，D 值为 0；第二期的第一个观测的 D 值为 $1/(1-\rho)$；其他第二期的观测的 D 值为 1。

第二，对于所有第一期的观测的乘积项，DX_t 的值为 0；第二期第一个观测的乘积项为 X_t；对于其他第二期的观测，其乘积项值为 $(X_t-\rho X_{t-1})$。

第7章

结　论

我们知道，恰当地运用虚拟变量可以大大提高回归模型的灵活性。然而，恰当地使用和解释虚拟变量涉及很多复杂的问题。本书的目的是通过分析逐步复杂的情况，为读者提供一些使用虚拟变量的指导。这些方法当然不完全，因为我们仅把视线局限于虚拟变量作为单方程模型的自变量中。但是，因子分析中的二元变量使用方法，如结构方程体系中的内生和外生变量，或者单方程或多方程系统中的因变量受到越来越多的关注。想追求更高深的定量数据分析的读者可以参考如下文献：Maddala(1983)处理回归模型中多线因变量的方法；Haberman(1978、1979)的两卷书；Goodman(1978)有关定量数据分析的文章；Aldrich 和 Nelson(1984)关于线性概率、logit 和普罗比模型概论；Allison(1984)有关事件历史模型的讨论；Muthen(1984)和 Shockey(1988)对无法观测变量的离散数据模型的讨论；Clogg 和 Goodman(1984、1985)的潜在结构分析以及 Winship 和 Mare(1983、1984)关于离散数据的结构方程模型和回归模型的文章。

注释

[1] 为了产生独立可靠的估计，就必须保证有足够数量的黑人。因此，我们对黑人区的住户进行过度采样。简单来说，在该测试中，我们用的是没有进行加权的数据。

[2] 实际意义较次要的是剩余的职业类别虚拟变量的相关系数。因为它们代表了所有不同且互斥的单一属性类别（如职业），这些虚拟变量都必须是逆向联系的，即相关性为负。在二分变量中，如同该例的种族，BLACK 和 WHITE 之间的相关性为－1.00。对于多分类变量，尽管相关性仍为负，但却不可能负得如此完美。如果一个受访者为服务业工作者，那他一定不是初级白领、技术工人、操作工人或者劳工，但是如果一个受访者不是服务业工作者，那他不一定是操作工人、劳工等等。两个虚拟变量之间的相关性大小是每两个变量中被编码为 1 的事件数和样本大小的函数。在表 2.3 中，技术工人和操作工人包含的事件最多，因此这两类的相关性最高，为－0.328。相反，服务业工作者和劳工是事件数最少的两类，其相关性仅为－0.108。在这种情况下，两个虚拟变量的相关性等于 $-[(p_j p'_j)/(1-p_j)(1-p'_j)]^{1/2}$。

[3] 我们必须记住，相关系数对变量的方差非常敏感。对于虚拟变量，相关系数由类别的相对频数而定。

[4] 由于模型 1 为一个二元回归模型，因此，F 检验和 t 检验是等价的，该检验的 t 值（－18.7）是 F 值（348.3）的平方根。从模型回归估计得到的信息和从单因素方差分析得到的结果一样，当相同的组均值估计出现时，F 检验会得出相同的结果（同样的数值）。另外，η^2（在该例中等于 0.09792）和模型 1 中的 R^2 也是一样的。

[5] F 检验的自由度包括了与回归平方和及残差平方和有关的自由度。回归平方和的自由度和模型中自变量的数目是相等的，在该情况下，如果模型包含 5 个虚拟变量，那么自由度为 5。残差平方和的自由度等于 $N-k-1$，其中 N 是观测数，k 是模型中自变量的个数。当 F 检验是由 R^2 和 $(1-R^2)$ 的比率计算出来的时候，自由度则如上所述。

[6] 从数学的角度来讲，参照组的选择是随意的，研究者可以选择不同的参照组再进行回归，让计算机程序来提供利益 t 检验。

[7] 方程 4.4 通过两个系数的加和来捕捉黑人在不同职业类别中的期望收入差异。

[8] 事实上，黑人中由显著职业差异导致的平均收入差异很少，该现象大部分归因于另外两个被控制了的变量——教育年限和工作任期。如果我

们所估计的模型只包含种族、职业和种族与职业的乘积项，那么回归结果应该为：

$$\begin{aligned}E(Y_i)=&10960.3-3958.4(\mathrm{BLACK})-2898.9(\mathrm{OCC}_2)-3625.6(\mathrm{OCC}_3)\\&-4875.0(\mathrm{OCC}_4)-6154.7(\mathrm{OCC}_5)-6182.9(\mathrm{OCC}_6)\\&+1747.8(\mathrm{BLOCC}_2)+1781.6(\mathrm{BLOCC}_3)+2594.5(\mathrm{BLOCC}_4)\\&+3238.6(\mathrm{BLOCC}_5)+2885.4(\mathrm{BLOCC}_6)\end{aligned}$$

其所有系数的 t 值都比±2.00 大。用方程 4.5 来构造黑人在不同职业中平均收入差异相当的 t 检验，我们发现，只有初级白领和高级白领的差异没有达到 0.05 的显著水平。读者也可以用以上估值来证明该模型设定中的均值和由种族内不同职业计算出来的均值（见表 2.2）是等价的。

[9] 教育年限和工作的差异效应与职业类别相关的问题是可能的。在该例中，我们对黑人和白人之间的差异估计感兴趣，不仅在于其期望收入的差异，还有和收入水平相关的结构影响问题。除了看种族之间的教育效应，我们还需在假设中加入职业类别的影响。在同一个雇主下多工作一年，其作为稳定性指标、作为对公司的忠诚或者特定的职业训练，可能技术工人相对于劳工会得到更多的收入。多一年的教育年限，专业人士可以挣得更高的收入，但这一点却不适用于工厂操作工人，因为这些不会增加他们职位提升的机会。如果我们的兴趣在于教育和工作任期的职业影响，乘积项可以根据每个职业虚拟变量和教育年限编码，也可以对每个职业虚拟变量和工作任期的乘积项进行编码。如果我们只检验职业差异效应并且发现教育和工作期限的差异影响非常显著，那么，就要构建 6 对平行线来看教育差异效应，每对实线都针对一个特定的职业组群，其斜率对于黑人和白人是一样的，但是对于不同的职业类别，斜率是不同的。关于描述工作任期差异效应的虚线，其趋势也一样。

[10] 反对一个零假设，即 $\beta=0$，对−0.16 进行 t 检验会产生一个−1.45 的 t 值；对−0.32 进行 t 检验，会产生一个−2.29 的 t 值。

[11] 有关两个种群参数的等效检验已经提出。其中一个检验适应于方差检验的框架（Chow, 1960; Maddala, 1992）。其步骤是我们先分别估计每个组群的回归模型，然后从不同的回归结果中获得其残差平方和（RSS_j）。并且，还需要对合并的样本进行回归估计，同时得到合并样本的残差平方和 RSS。对于该参数等效的 F 检验（有 $k+1$ 和 n_1+n_2-2 个自由度），$F=[(\mathrm{RSS}_{\mathrm{pooled}}-\sum\mathrm{RSS}_j)/(k+1)]/[\sum\mathrm{RSS}_j/(n_1+n_2-$

$2k-2)$],其中,$\sum \text{RSS}_j$ 为对不同组群进行回归估计得到的 RSS 之和,k 为模型中自变量的个数,n_1 和 n_2 分别为两个组群的观测数。该例的 F 值已经大到可以拒绝零假设(参数的平等性),这说明在两组群中,不是所有的自变量的影响都一样。然而,该检验并不能表明哪些参数是不同的。

[12] 残差平方和 $\text{RSS}/(n-k-1)$ 中,n 为事件数,k 为模型中的自变量个数,RSS 是残差平方和,可由 $\sum e_i^2$ 算得,但是一般的统计软件程序都会提供该结果。在 OLS 假设下,该数通常可以提供对 σ^2 的无偏估计,即总体的方差 u_i。

[13] 在计算对数时,我们必须说明其基数。最普遍的基数就是 e,其通常也被称为"自然对数",还有一个就是 10。e 的值约为 2.72。要对以 e 为底的 X 取对数,我们需要确定 X 所需要的 e 的幂次。同样,如果取以 10 为底的 X 的对数,我们就需要确定 X 所需要的 10 的幂次是多少。对数模式会使分布变得与之前不同,其非线性也非常明显,因为以 10 为底的对数值 1、2、3 对所对应的最初分布数值为 10、100 和 1000。

[14] 估计黑人在不同职业内影响的显著性与确定黑人和白人在不同职业内差异的显著性的过程是一样的。黑人在高级白领中的效应可以通过 BLACK 回归系数捕捉到,在初级白领中的效应可以看 B_1+B_9;对于技术工人,该效应为 B_1+B_{11};对于服务业工作者,为 B_1+B_{12};对于劳工,黑人的效应为 B_1+B_{13};如果要看种族效应在职业类别内是否显著,即看其相对其他特定效应的净效应,我们要用方程 4.5 的 t 检验来表达。5 个职业类别的 t 值分别为-4.07、-5.58、-5.63、-1.90和-4.70。

[15] 当创建对比编码变量时,我们有很多选择。比如,可以用技术工人与操作工人、服务业工作者和劳工的组合来进行比较。这样,技术工人则被编码为 1,而操作工人、服务业工作者和劳工就被编码为$-1/3$。之后,再比较操作工人与服务业工作者和劳工的组合,这样,最后一个比较也和文中所述的一样了。

[16] 熟悉传统方差检验的读者可能知道,这种情况和同组群大小要求的情况的相似处在于,它们都是在一个 n 因素方差分析中作出正交设计的。

[17] 由 C_5 定义的对比显著性的替换性检验,在第 3 章描述关于服务业工作者和劳工之间的差异时已被提及。方程 3.1 为包含二进制编码虚拟变量的两组群间的差异提供了检验方程。读者可以用表 3.1 中的结果来证实两个过程的 t 值相同,而且其对比和标准差都与方程 3.1 中定义的均值差异和标准差差异估计一样。

参考文献

Alba, R. D. (1988)"Interpreting the parameters of log-linear models," in J. S. Long(ed.) *Common Problems/Proper Solutions: Avoiding Error in Quantitative Research*. Newbury Park, CA: Sage.

Aldrich, J. H., and Nelson, F. D. (1984) *Linear Probability, Logit, and Probit Models*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—045. Beverly Hills, CA: Sage.

Allison, P. D. (1984) *Event History Analysis: Regression for Longitudinal Event Data*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—046. Beverly Hills, CA: Sage.

Amemiya, T. (1986) *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Bartlett, M. S. (1937) "Properties of sufficiency and statistical tests." *Proceedings of the Royal Society of London 160* (Series A).

Berry, W. D., and Feldman, S. (1985) *Multiple Regression in Practice*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—050. Beverly Hills, CA: Sage.

Bohrnstedt, G., and Knoke, D. (1982) *Statistics for Social Data Analysis*. Itasca, IL: F. E. Peacock.

Chen, C. (1984) "The structural stability of the market model after the Three Mile Island accident." *Journal of Economics and Business* 36: 133—140.

Chow, G. C. (1960) "Test of equality between sets of coefficients in two linear regressions." *Econometrica* 28:591—605.

Clogg, C. C., and Goodman, L. A. (1984) "Latent Structure analysis of a set of multidimensional contingency tables." *Journal of the American Statistical Association* 79:762—771.

Clogg, C. C., and Goodman, L. A. (1985) "Simultaneous latent structure analysis in several groups," in N. Tuma (ed.) *Sociological Methodology*. San Francisco: Jossey-Bass.

Cohen, J., and Cohen, P. (1983) *Applied Multiple Regression* (2nd *ed.*). Hillsdale, NJ: Lawrence Erlbaum.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981) "A comparative study of tests for homogeneity of variances with applications to the

Outer Continental Shelf bidding data." *Technometrics* 23:351—361.

Darlington, R. B. (1990) *Regression and Linear Models*. New York: McGraw-Hill.

Dunn, O. J. (1961) "Multiple comparisons among means." *Journal of the American Statistical Association* 56:52—64.

Glejser, H. (1969) "A new test for homoscedasticity." *Journal of the American Statistical Association* 64:316—323.

Goldfeld, S. M., and Quandt, R. E. (1972) *Nonlinear Methods in Econometrics*. Amsterdam: North-Holland.

Goldfeld, S. M., and Quandt, R. E. (1978) "Asymptotic tests for the constancy of regressions in the heteroscedastic case." *Research Memorandum* No. 229, Econometric Research Program, Princeton University.

Goodman, L. A. (1978) *Analyzing Qualitative/Categorical Data*. Cambridge, MA: Abt Associates.

Gujarati, D. N. (1970) "Use of dummy variables in testing for equality of sets of coefficients in two linear regressions: A note." *American Statistician* (February).

Gujarati, D. N. (1988) *Basic Ecnomiteirsc* (2nd *ed.*). New York: McGraw-Hill.

Haberman, S. J. (1978) *Analysis of Qualitative Data*, Vol. 1: Introductory Topics. New York: Academic Press.

Haberman, S. J. (1979) *Analysis of Qualitative Data*, Vol. 2. New York: Academic Press.

Halvorsen, R., and Palmquist, R. (1980) "Interpretation of dummy variables in semilogarithmic equations." *American Economic Review* 70: 474—475.

Idler, E. L., and Kasl, S. (1991) "Health perceptions and survival: Do global evaluations of health status realy predict mortality?" *Journal of Gerontology* 46(2):S55—65.

Jaccard, J., Turrisi, R., and Wan, C. K. (1990) *Interaction Effects in Multiple Regression*. Sage University Paper series on Quantitative Applicatons in the Social Sciences, 07—072. Newbury Park, CA: Sage.

Johnston, J. (1984) *Econometric Methods* (3rd *ed.*). New York: McGraw-Hill.

Kendall, M. G., and Stuart, A. (1979) *The Advanced Theory of Statistics*, Vol. 2(4th ed.). New York: Charles Griffin.

Kmenta, J. (1986) *Elements of Ecnonomtrics* (2nd *ed.*). New York: Macmillan.

Levene, H. (1960) "Robust tests for equality of variances," in I. Olkin

(ed.) *Contributions to Probability and Statistics*. Stanford, CA: Stanford University Press.

Lewis-Beck, M. S. (1980) *Applied Regression: An Introduction*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—022. Beverly Hills, CA: Sage.

Long, J. S., and Miethe, T. D. (1988) "The statistical comparison of groups," in J. S. Long (ed.) *Common Problems/Proper Solutions: Avoiding Error in Quantitative Research*. Newbury Park, CA: Sage.

Maddala, G. S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Maddala, G. S. (1992) *Introduction to Econometrics* (2nd *ed.*). New York: Macmillan.

Miller, R. G., Jr. (1966) *Simultaneous Statistical Inference*. New York: McGraw-Hill.

Muthen, B. (1984) "A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators." *Psychometrics* 46:115—132.

Ryan, T. A. (1960) "Significance tests for multiple comparisons of proportions, variances, and other statistics." *Psychological Bulletin* 57: 318—328.

Schroeder, L. D., Sjoquist, P. L., and Stephan, P. E. (1986) *Understanding Regresion Analysis: An Introductory Guide*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—057. Beverly Hills, CA: Sage.

Shockey, J. W. (1988) "Latent-class analysis: An introduction to discrete data models with unobserved variables," in J. S. Long (ed.) *Common Problems/Proper Solutions Avoiding Error in Quantitative Research*. Newbury Park, CA: Sage.

Suits, D. (1983) "Dummy variables: Mechanics v. interpretation." *Review of Ecnomics and Statistics* 66:177—180.

Welch, B. L. (1938) "The significance of the difference between two means when the poplation variance are unequal." *Biometrika* 29:350—362.

Winship, C., and Mare, R. D. (1983) "Strutural equations and path analysis for discrete data." *American Journal of Sociology* 89:54—110.

Winship, C., and Mare, R. D. (1984) "Regression models with ordinal variables." *American Sociological Review* 49:512—525.

译名对照表

additive effect	相加效应
bias	偏差
binary coding	二进制编码
bivariate regression	二元回归
chance fluctuation	随机波动
contrast coding	对比编码
correlation	相关性
covariance	协方差
cross sectional data	截面数据
cross-tabular analysis	列联表分析
curvilinearity	曲线性
dependent variable	因变量
descriptive Statistics	描述性统计
differential effects	差异效应
discrete variable	离散变量
dummy variable	虚拟变量
effects coding	效果编码
expected value	期望值
explanatory power	解释功效
Fisher's protected t method	Fisher 的 t 保护方法
fit of the model	模型拟合
homogeneity	同质性
hypothetical data	假设数据
independent variable	自变量
incremental F test	增量 F 检验
interaction	交互作用
interaction effect	交互效应
interaction term	交互项
intercept	截距
magnitude	强度
marginal value	边际值

mean regression sums of squares	平均回归平方和
measures of association	相联度量
midrange mean value	中距均值
moderating effect	调解效应
multiplicative effect	乘积效果
nominal variable	名义变量
non-independent test	非独立检验
non-normality	非常态性
null hypothesis	零假设
one-way analysis of variance	单向方差分析
orthogonal	正交
partial correlation	偏相关
partial effect	局部效应/偏效应
partial slope	偏斜率
partialing procedure	剔除过程
piecewise linear regression	分段线性回归
point biserial correlation coefficient	点二列相关系数
point estimate	点估计
polynomial regression	多项式回归
polytomous variable	多分类变量
population	整体
pooled estimate	合并估计值
power	检定力
reference group	参照组
regressor	回归因子
regression coefficient	回归系数
residual sum of square	残差平方和
robustness	稳健性
sampling error	抽样误差
semilogarithmic equation	半对数方程
simultaneous statistical inference	同时统计学推论
single explanatory characteristic	单解释性特征
slope	斜率

statistically significant	显著性水平
t statistic	*t* 统计量
test statistic	检验统计量
time-series data	时间序列数据
two-tailed test	双尾检验
variance	方差
variation	变异
weight	加权



格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用logistic回归分析（第二版）
14. logit与probit：次序模型和多类别模型
15. 定序因变量的logistic回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic回归入门
39. 解释概率模型：Logit、Probit以及其他广义线性模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL导论
43. 非参数回归：平滑散点图
44. 广义线性模型：一种统一的方法
45. Logistic回归中的交互效应
46. 应用回归导论
47. 档案数据处理：研究“人生”
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践


微信


天猫

上架建议：社会研究方法


ISBN 978-7-5432-2642
9 787543 226425

定价：25.00元
易文网：www.ewen.co
格致网：www.hibooks.cn

[General Information]
书名=虚拟变量回归
作者=（美）梅丽莎?A.海蒂著
页数=128
SS号=14074634
DX号=
出版日期=2016.08
出版社=格致出版社；上海人民出版社